Jornadas de Seguimiento de Proyectos, 2010
Programa Nacional de Tecnologías Informáticas

Probabilistic Graphical Models in Supervised Multi-dimensional Classification Problems TIN2008-06815-C01/C02

Jose A. Lozano Alonso *
Intelligent Systems Group
Department of Computer Science and Artificial Intelligence
The University of the Basque Country

Luis Baumela Molina †
Departamento de Inteligencia Artificial
Universidad Politécnica de Madrid

Abstract

This research project deals with problems in the field of supervised classification where there are several variables to be classified (we will use the term multi-dimensional classification). Probabilistic graphical models based on directed acyclic graphs are the chosen formalism to solve this problem. We plan to carry out new methodological developments able to solve the special characteristics of having more than one variable to be predicted. Particularly we will develop new scores and probabilistic model learning algorithms. We will also tackle the problem of feature subset selection in multi-dimensional classification, and design new classification performance methodologies for the problem at hand. The developed methodology will be put in practice by solving real problems in the fields of bioinformatics and computer vision.
Keywords: supervised classification, probabilistic graphical models, multi-dimensional classification, bioinformatics, computer vision

1 Project Goals, Schedule and Research Team

1.1 Goals

This project tries to develop methodological and computational contributions to the problem of predicting a set of class-variables by means of the Bayesian network paradigm. We will briefly summarize these contributions in the next lines:

*Email: ja.lozano@ehu.es
†Email: lbaumela@fi.upm.es

- Development of learning algorithms for supervised Bayesian classifiers in multi-dimensional classification ("outputs").

  The learning process is conducted in a reduced search space with two different layers for domain variables. While a first layer is used for the class-variables, predictive variables are located in the second layer: dependence arcs are not allowed from this second level to the first one. Probabilistic relationships of different nature have to be learned from data: relations between the variables to be predicted, relations between the predictive variables, and the arcs from class-variables to predictors. A "score + search" approach will be used to deal with the learning process, and the complexity degree of the relationships between domain variables will be fixed in advance. "Filter" (penalized likelihood; regularized likelihood and "wrapper"; learning with different misclassification costs scores) approaches will be used. In fact one of the main challenges of this project is the generalization of these well-known scores to supervised problems with more than one variable to be predicted. As a search has to be performed in the space of possible models, optimization search heuristics of different complexity will be used, ranging from local techniques to randomized, population-based procedures such as estimation of distribution algorithms.

- Development of feature selection algorithms for problems with more than one class-variable to be predicted.

  Based on the multi-dimensional nature of the output to be predicted, the generalization of filter feature selection algorithms to this kind of problems will be one of the objective of the project. The filter techniques to develop will be univariate (evaluation of the correlation of each predictive variable with respect to the vector of classes) and multivariate (evaluation of the correlation of a subset of predictive variables with respect to the vector of classes).

- Development of accuracy evaluation and comparison techniques for multi-dimensional class models.

  In order to develop the previously exposed wrapper learning approaches, the generalization of Brier score and ROC curve evaluation techniques to this kind of problems with various classes to be predicted will be mandatory. Based on the "multiple response permutation procedure" paradigm, our aim is also to generalize the expression of statistical hypothesis tests for the comparison of a group of classifiers. This extension will be done when the comparison is performed in a single dataset and in a collection of datasets.

- Application in bioinformatic problems.

  We plan to apply the algorithms developed with the objective of obtaining good solutions to the problem of the prediction of the prognosis variables in inflammatory bowel diseases. Particularly we will work with Chron disease, ulcerative colitis and irritable bowel syndrome. The objectives are to increase the prediction ability of current models by the used of multi-dimensional classification models. We will also use features subset selection techniques in order to select those variables that are more relevant to the problem at hand and therefore to obtain information about the disease from the selected variables.

- Application in computer vision problems.

We work on the problem of demographic classification based on the appearance of the face. We will assume that age and ethnicity are continuous variables and prove that, by considering the multi-dimensional relations among the demographic variables, we will be able to build better classification systems.

1.2 Schedule

First of all we would like to point out that this project, on the contrary of most of the projects presented in this meeting, is a two-year project, and therefore this report refers only to one year of work.

The tasks in which the project is mainly divided and its time schedule are described below:

1. Literature review (M(onth)1-M(onth)24)
2. Development and implementation of classification evaluation methods (M1-M12)
3. Development and implementation of scores (M5-M14)
4. Development and implementation of search algorithms (M5-M14)
5. Comparison between classifiers (M12-M18)
6. Application in the bioinformatics field (M9-M24)
7. Application in the Computer Vision field (M1-M24)

1.3 Project Team

This a "coordinated project" and it means that the project is carried out by several research teams. In this particular case there are two research teams, one at The University of the Basque Country (UPV–EHU) and the other at "Technical University of Madrid" (UPM). The research team at the UPV–EHU is composed of four senior researchers (more than 5 years of research experience after the PhD), four post-doc researchers and four PhD students. In the case of the group of UPM this is composed of three senior researchers, one post-doc, and three PhD students.

2 Success level reached in the project

In order to properly organise this section we have divided it into several sections, each of one corresponds with one of the main objectives of the project.

2.1 Development of learning algorithms

In this objective we have reached a high level of accomplishment. We have developed several approaches to learn multi-dimensional Bayesian network classifiers.

- A first approach is based on multiobjective optimization. Given that in a multi-dimensional supervised classification problem there are several variables to classify, it is possible to separately assign to each class variable a performance measure (in our case we have used accuracy). Therefore, we can think in the process of learning a Bayesian network classifier as the search for a classifier that maximizes the accuracy of each variable separately. However, it is not difficult to check that in most of the domains there does not exist such a classifier. Even more it usually happens that if a classifier B improves the accuracy of another classifier B' in class variable C_i then B deteriorates the accuracy of other variable C_j in relation with B'. Therefore this environment suggests a multiobjective approach: The goal is to find classifiers in such a way that if another classifier improves the accuracy of one of the class variables it implies a deterioration in the accuracy of other variable.

  In order to apply this multiobjective approach we have carried out a search in the space of Bayesian network classifiers. This space is composed of all the classifiers whose structures are as follows: We consider two different layers for domain variables. While a first layer is used for the class-variables, predictive variables are located in the second layer: dependence arcs are not allowed from this second level to the first one. We allow three kind of relations: (i) between the variables to be predicted, (ii) relations between the predictive variables, and (iii) arcs from class variables to predictors. The algorithm to carry out the search in this space is a evolutionary algorithm for multiobjective optimization called MOEA/D. In order to evaluate the whole approach we have used synthetic data as well as real data coming from the UCI repository. A paper describing this research is under submission in the journal *Machine Learning* (Rodríguez et al 2010b).

- In a second approach we have developed learning algorithms for Bayesian networks classifiers in multi-dimensional semisupervised problems. In semisupervised problems we have labeled and unlabeled data and the objective is to find a classifier that can take advantage of the use of unlabeled data. For single class problems it has been proved that the classifier with the right model (the one that generated the data) improves accuracy by adding the unlabeled data, however in the case of using the wrong model the addition of unlabeled data to learn the model can hurt the performance. This implies that a learning algorithm for semisupervised problems needs to carry out a search in the space of models. In this case we have developed a wrapper approach to learn multi-dimensional Bayesian networks classifiers. Particularly the algorithm moves in the same classifiers space as in the previous approach. In this case we have developed a search algorithm that initially learns an structure between the classification variables and once this structure is fixed it learns the arcs between the predicitive variables and also between the classification variables and the predictive variables. In the whole process the EM algorithm is used to deal with the unlabeled data. All this work is almost finished and is going to be submitted to an special issue of the *International Journal of Electronic Commerce* (dealine is April 15th) where an extended abstract has been initially accepted.

- In a third approach, the problem is tackled by a general family of models, called multi-dimensional Bayesian network classifiers (MBCs). This probabilistic graphical model organizes class and feature variables as three different subgraphs: class subgraph, feature subgraph, and bridge (from class to features) subgraph. Under the standard 0-1

loss function, the most probable explanation (MPE) must be computed, for which we provide theoretical results in both general MBCs and in MBCs decomposable into maximal connected components. Moreover, when computing the MPE, the vector of class values is covered by following a special ordering (gray code). Under other loss functions defined in accordance with a decomposable structure, we derive theoretical results on how to minimize the expected loss. Besides these inference issues, we introduce flexible algorithms for learning MBC structures from data based on filter, wrapper and hybrid approaches. The cardinality of the search space is also given. Experimental results with three benchmark data sets are encouraging, and they outperform state-of-the-art algorithms for multi-label classification. The manuscript has been published as a technical report at the Department of Artificial Intelligence (UPM) (Bielza et al. 2010c) and also submitted to the *Journal of Machine Learning Research Journal* (Bielza et al. 2010b).

- In a fourth approach we aim at solving the learning problem by means of regularization methods. The use of these methods will provide a more stable and robust approaches, specially for situations with sparse data sets. Nowadays we are developing an approach that takes into account at the same time the regularization of the structure of the graphical model (Vidaurre et al. 2010a) –in press in the *IEEE Transactions on Systems, Man, and Cybernetics-Part B*– and the parameters (Bielza et al. 2009)–published in *Methods of Information in Medicine*.

- The use of metaheuristic based search is fundamental to move in an intelligent way in the space of multi-dimensional Bayesian network classifiers or in its corresponding ordering space. Researchers from both group are familiarized with genetic algorithms as metaheuristics, and we plan to use them in the near future for searching the optimal MBC. The results obtained in our recent work (Bielza et al. 2010a) could be applied. Also in a collaboration between both research groups we have developed a Matlab package for the implementation and analysis of estimation of distribution algorithms (Santana et al. 2010) that will be published in the *Journal of Statistical Software*.

- The combination of regularization and nearest neighbor has been explored, for regression problems, by members of the researcher group (Vidaurre et al. 2010b) in a work submitted to *Statistics and Computing* journal. This piece of work will allow us to extend the scope of the project from multi-dimensional supervised classification problems to multi-dimensional regression ones.

2.2 Development of feature selection algorithms

In this goal we have gone beyond the development of feature selection algorithms for supervised multi-dimensional classification problems and we have dealt with preprocessing methods for supervised multi-dimensional classification problems that include: techniques to deal with missing values, discretization techniques, and also feature subset selection algorithms. In all the three problems (missing processing, discretization, and feature selection) we have implemented two basic approaches taking into account the techniques for single variable classification: i) to consider only one variable that is the Cartesian product of all the variables, and (ii) to mixed the results obtained when applied single variable method to each variable separately. In addition to adapt single variable methods to multi-dimensional classification we have also created new

methods departing from them. Particularly we have developed a new method for dealing with missing data based on the classical "Cmean" techniques. In the case of discretization we have concentrated in adapting the classical "Fayyad & Irani" method to the field of supervised multi-dimensional classification. Finally in the case of feature selection an extension of the classical "Correlation-based Feature Selection (CFS) has been designed. All this work is under revision on the journal *Applied Soft Computing* (Fernandes et al 2010).

A possible extension of the standard multi-dimensional classification problem consists in situations where some of the class variables contain positive and unlabelled examples. For these difficult situations we have developed a preliminary work that only considers one class variable where one adaptation of the CFS technique has been proposed. See the paper by Calvo et al. (2009), published in *Pattern Recognition Letters*, for details.

2.3 Development of accuracy evaluation and comparison techniques for multi-dimensional class models

New performance evaluation metrics adapted from the single-class setting have been introduced. Between them the mean accuracy over the d class variables, the global accuracy over the d-dimensional class variable, the micro $F1$, and the macro $F1$. See Bielza et al (2010b), Bielza et al (2010c) and Rodríguez et al (2010b) for details.

In addition to the introduction of the previously mentioned performance measures, a theoretical study has been carried out for the most commonly used error estimation technique: k-fold crossvalidation (Rodríguez et al, 2010a). In this study the variance of k-fold that depends on the fold and the one that depends on the sample is estimated. This is a first step to carry out the same analysis in multi-dimensional classification.

2.4 Application in bioinformatic problems

We have just finished with the approaches to learn multi-dimensional Bayesian network classifiers and plan to deal with the bioinformatics problems in the second year of the project. However we would like to emphasize the work we have done in other two application fields. Firstly we have applied some of the developed techniques in the recruitment prediction of several fish species (Fernandes et al, 2009a; Fernandes et al, 2009b) and we also plan to use the new developed preprocessing techniques to get better prediction models in this area. A second application field is sentiment analysis. Particularly the developed search algorithm for semisupervised learning has been applied to deal with real data coming from the product "asomo" of the company "socialware", that is dedicated to mine blogs.

Nowadays, publishers of scientific journals face the tough task of selecting high-quality articles that will attract as many readers as possible from a pool of articles. The possibility of a journal having a tool capable of predicting the citation count of an article within the first years after publication would pave the way for new assessment systems. In a recent paper (Ibañez et al. (2009)) we have presented a new approach based on building several prediction models for the *Bioinformatics* journal. These model predict the citation count of an article within 4 years after publication. To build these models, tokens found in the abstracts of Bioinformatics papers have been used as predictive features, along with other features like the journal sections and the number of two-week post-publication periods.

		UPV-EHU	UPM	Project
Book Chapters	International	3	2	3
Journals	International	11	11	15
	National		1	1
Conferences	International	5	2	6
	National	2		2
Technical reports	National	4	2	5

Table 1: Summary of the publications

2.5 Application in computer vision problems

The first step in a demographic classification system is detecting, and tracking the face. We have studied both 2D and 3D model-based approaches for face tracking. In the 2D approaches we assume that the face is a planar textured target. We have developed efficient techniques to track a deforming 2D target under strong illumination changes (Buenaposada et al. (2009)). In the 3D approach we model the face as a 3D textured mesh that can deform based on a set of linear modes of deformation (3D Morphable Model). We have developed a direct approach for efficiently fitting a 3D Morphable model to a moving and deforming face (Muñoz et al. (2009)). Finally, we have also studied the application of snakes to track a model-free deforming target. In this context, we have introduced a novel curve evolution procedure based on morphological operators which is faster and more stable than traditional explicit schemes (Álvarez et al. 2010).

We have also investigated the construction of a demographic classification system. We have adopted an appearance-based approach considering the face texture as a vector of predictive variables. We have developed statistical classifiers for demographic classification based on linear dimensionality reduction (Bekios et al. (2010a)) and on class-conditional probabilistic principal component analysis (Bekios et al. (2010b)).

3 Summary of the results

In this section we report on the various results we can consider as a consequence of the developed work. We would like to point out that there has been an extensive collaboration between both, UPV-EHU and UPM research groups and therefore we have written only one joint section. We will make a distinction of the results of each group for some of the contributions.

A first set of results we can take into account is the obtained publications. The global numbers in terms of them can be consulted in Table 1. Some of these publications have been cited along this document and can be consulted in the reference section. It is important to take into account that the project budget supports the general research of the group, and therefore most of the work done in the group references the project.

In addition to the publication record we would like to emphasize the research training abilities of the research groups. This is captured by the fact that two PhD dissertation have been presented by members of the project. Furthermore one of these thesis has been European PhD awarded.

In the following a list of PhD dissertations co-supervised by members of both research groups follows:

- D. A. Morales (2008). *Clasificadores Bayesianos en la Selección Embrionaria en Tratamientos de Reproducción Asistida.* Departamento de Ciencias de la Computación e Inteligencia Artificial (UPV–EHU). December 2008. Supervised by P. Larrañaga and E. Bengoetxea.
- R. Armañanzas (2009). *Consensus Policies to Solve Bioinformatics Problems through Bayesian Networks Classifiers and Estimation of Distribution Algorithms.* Departamento de Ciencias de la Computación e Inteligencia Artificial (UPV–EHU). April 2009. Supervised by P. Larrañaga and I. Inza. European PhD awarded.

The following two PhD dissertations also co-supervised by members of the two research groups have been deposited in January 2010 and will be presented in the next months:

- A. Pérez (2010). *Supervised Classification in Continuous Domains with Bayesian Networks.* Departamento de Ciencias de la Computación e Inteligencia Artificial (UPV–EHU). Supervised by P. Larrañaga and I. Inza. European PhD awarded.
- T. Miquélez (2010). *Avances en Algoritmos de Estimación de Distribuciones. Alternativas en el Aprendizaje y Representación de Problemas.* Departamento de Arquitectura y Tecnología de Computadores (UPV–EHU). Supervised by P. Larrañaga and E. Bengoetxea.

Another point that deserves interest is the set up of relations with other research groups at all levels, national and international. In this respect we would like to remark the new groups that we have established relation with. Particularly, UPV-EHU group has got a European grant under the Seventh Framework Programme titled "Natured Inspired Computation and Its Applications" in collaboration with The University of Birmingham, Leiden University and three Chinese universities. In addition two more proposal have been submitted with:

- CINVESTAV-IPN, MEXICO (AECID)
- Seoul National University, Korea (Korea Science Foundation)

Members of the UPM research group have been invited to give some talks in different departments, conferences and scientific meetings. Here is a list containing some of them:

- P. Larrañaga, C. Bielza (2009). *Estimation of Distribution Algorithms and Regularization.* University of Essex
- P. Larrañaga, C. Bielza (2009). *Computational Intelligence for Neuroscience.* Tutorial at the Discovery Science and Algorithmic Learning Theory Conference. Porto.
- P. Larrañaga, C. Bielza (2009). *Machine Learning and Neuroscience.* Aveiro University
- P. Larrañaga (2010). *Probabilistic Graphical Models and Evolutionary Computation.* Plenary speaker at the 2010 IEEE World Congress on Computational Intelligence. Barcelona
- P. Larrañaga (2010). *Bayesian Networks and Evolutionary Computation.* Plenary speaker at the Argentine Symposium on Artificial Intelligence. Buenos Aires

- P. Larrañaga (2010). *Multilabel Classification.* Invited speaker at the Simposium TAMIDA de Minería de Datos del CEDI 2010. Valencia

- P. Larrañaga (2010). *Probabilistic Graphical Models for Multidimensional Classification of Data Streams.* Indian Institute of Science of Bangalore

- P. Larrañaga (2010). *Estimation of Distribution Algorithms.* National University of Seoul

4 Conclusions

Globally we think that most of the objectives proposed in the project have already reached a high level of accomplishment. However there are points where we have to put more effort in the second year left.

We are particularly happy with the level of achievement of the **i)**, **iii)** and **v)** objectives. Although, of course, more methodological developments can be carried out in the future in those sections. On the other hand the objectives that have received lower level of achievement than the other are the second and the fourth. These are the points in which we will to put the most effort in the second year of the project.

Taking into account other aspects of the project, we plan for the future to increase our relation with companies and to consider the possibility of involving ourselves in another European project.

References

[1] L. Álvarez, L. Baumela, P. Márquez (2010). Morphological Snakes. *International Conference on Computer Vision and Pattern Recognition*, submitted.

[2] J. Bekios, J.M. Buenaposada, L. Baumela (2010). Revisiting linear discriminant techniques in gender recognition. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, submitted.

[3] J. Bekios, J.M. Buenaposada, L. Baumela (2010). Gender recognition by Class-Conditional Probabilistic Principal Component Analysis. *Revista Computación y Sistemas*, (invited submission).

[4] 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.

[5] C. Bielza, J.A. Fernández del Pozo, P. Larrañaga, E. Bengoetxea (2010a). Multidimensioanl statistical analysis of the parameterization of a genetic algorithm for the optimal ordering of tables. *Expert Systems with Applications*, 37, 804-815.

[6] C. Bielza, G. Li, P. Larrañaga (2010b). Multi-dimensional classification with Bayesian networks. *Journal of Machine Learning Research*, submitted.

[7] C. Bielza, G. Li, P. Larrañaga (2010c). Multi-dimensional classification with Bayesian networks. *Technical Report. Department of Artificial Intelligence. Technical University of Madrid. UPM-DIA-2010-01.*

[8] J.M. Buenaposada, E. Muñoz, L. Baumela (2009). Efficient illumination independent appearance-based face tracking. *Image and vision computing*, 27(5), 560-578.

[9] B. Calvo, P. Larrañaga, J.A. Lozano (2009). Feature subset selection from positive and unlabelled examples. *Pattern Recognition Letters*, 30, 1027-1036.

[10] J. A. Fernandes, X. Irigoyen, J. A. Lozano, I. Inza (2009a) Optimizing the number of classes in automated zooplankton classification. *Journal of Plankton Research*, 31(1), 19-29.

[11] J. A. Fernandes, X. Irigoien, N. Goikoetxea, J. A. Lozano, I. Inza, A. Perez and A. Bode (2009b). Fish recruitment prediction, using robust supervised classification methods. *Ecological Modelling*, 221(2), 338-352.

[12] J.A. Fernandes, X. Irigoien, J.A. Lozano, I. Inza, A. Pérez, J.D. Rodríguez (2010) Supervised pre-processing approaches in multiple class-variables classification for fish recruitment forecasting. *Applied Soft Computing*, submitted.

[13] A. Ibañez, P. Larrañaga, C. Bielza (2009). Predicting citation count of Bioinformatics apapers within four years of publication. *Bioinformatics*, Vol. 25, No. 24, 3303-3309.

[14] E. Muñoz, J.M. Buenaposada, L. Baumela (2009). A direct approach for efficiently tracking with 3D Morphable Models. *International Conference on Computer Vision,* 1615-1622.

[15] J.D. Rodríguez, A. Pérez and J.A. Lozano (2010a). Sensitivity Analysis of k-fold cross-validation in prediction error estimation. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 32(3), 569-574.

[16] J.D. Rodríguez, J.A. Lozano (2010b). Learning Bayesian netowrk classifiers for multi-dimensional supervised classification problems by means of a multi-objective approach. *Machine Learning*, submitted.

[17] D. Vidaurre, C. Bielza, P. Larrañaga (2010a). Learning an L1-regularized Gaussian Bayesian network in the equivalent class space. *IEEE Transactions on Systems, Man, and Cybernetics-Part B.* In press.

[18] D. Vidaurre, C. Bielza, P. Larrañaga (2010b). Lazy lasso for local regression. *Statistics and Computing*, submitted.

[19] R. Santana, C. Bielza, P. Larrañaga, J.A. Lozano, C. Echegoyen, A. Mendiburu, R. Armañanzas, S. Shakya (2010). MATEDA: A Matlab package for the implementation and analysis of estimation of distribution algorithms. *Journal of Statistical Software*, In press.