Jornada de Seguimiento de Proyectos, 2004
Programa Nacional de Tecnologías Informáticas

Modelado Individualizado de Secuencias Simbólicas (MOISES) TIC2002-04019-C03

Elvira Mayordomo Cámara [*]
Universidad de Zaragoza

Rafael Morales Bueno [†]
Universidad de Málaga

Glyn Morrill [‡]
Universidad Politécnica de Cataluña

Abstract

This project approaches the study, theoretical development, implementation, and empirical validation of formal concepts and the criteria for its use in the construction of predictive and descriptive models for symbolic sequences.

In particular we will study the following model types: compression mechanisms (finite-state compressors and Lempel-Ziv algorithms), generalizations of graphs (probabilistic extensions, decision trees, integration of decision trees with Markov hidden models), grammatical models (such as categorial grammars), time series, study of subsequences (discovery of episodes, frequent sets and association rules; learning of patterns of behaviour).

The empirical validation of models will be done through great amounts of real data: university students, oncological data from the Hospital Clínico Universitario, biological sequences, climatology data.

Keywords: symbolic sequences, compression algorithms, automatic learning, decision trees, data-mining, time series, categorial grammars, Turing-machine, Kolmogorov complexity

1 Project objectives

The abstraction of sequences of events (activities) over time are the sequences of symbols. We face here the problem of obtaining useful information from massive quantities of data where there is a sequentiality notion. Our global objective is the following. The study, the theoretical development, the implementation and the empirical validation of formal concepts and criteria for their use in the construction of descriptive and predictive models for symbolical sequences.

We will select families of possible models for sequence modelling and, for each model, we will define in which sense (or senses) it can be used to describe or predict a sequence.

[*] Email: elvira@unizar.es
[†] Email: morales@lcc.uma.es
[‡] Email: morrill@lsi.upc.es

Then the theoretical development will concentrate on the mathematical study of the modelling quality and the related problems that appear in this development. We next apply this model to particular sequences, looking for empirical validation of the mathematical study and for intuitive suggestions of new properties.

The work plan will consist roughly in three steps:

- Development of the formalisms.
- Programming (implementation of the formalisms studied in the first step, together with the collection of large amounts of real data).
- Applications.

The first step will have its core in the first year, but will continue in the second and third year with the revisions originated by implementations and applications. The second step will start in the second year.

We will focus on six different mode types. We next list the particular objectives that correspond to each of them.

Note. The project FRESCO, CICYT PB98-0937-C04, was performed from 2000 to 2002 and will be mentioned below.

1.1 Compression mechanisms

The relationship among compression, prediction methods and Hausdorff dimension is known in several particular contexts [32, 31, 26].

Our first objective is to extend this relationship in the broadest possible sense, both in terms of the resource-bounds imposed on the algorithm and in the sense of the performance measure used. In particular we are interested in the case of polynomial time compression algorithms such as Lempel-Ziv's, as well as in finishing the ongoing research on finite-state compressors. We also will explore further the role of Kolmogorov complexity in this compression vs. prediction approach.

Next we will use the obtained results in identifying the formal properties that characterize sequences that are very compressible by efficient compression mechanisms. Finally from this knowledge we expect to implement and validate the methods that allow an improvement in the compression ratio of interesting sequences.

1.2 Generalizations of graphs

We want to find new algorithmic schemes for the basic problems of construction and isomorphy of different families of graphs.

Our next objective is to consolidate the results on probabilistic (by sampling) construction of decision trees in TDIDT methods (obtained in the FRESCO project) by incorporating time. In particular we will

- incorporate techniques of prediction by voting
- create “forgetful” sampling algorithms, that won't keep the seen experiences but will keep the good properties of the constructed tree.

The third objective here is to integrate decision trees with Markov hidden models

- consolidating the results on suffix graphs of multiatribute prediction (MPSG) from the FRESCO project
- continuing the development of the morphological analyser and including new texts in the training process.

A more ambitious objective will be to understand the role of Galois connection as a unifying role of several of our approaches, in particular of graph based models and the discovery of episodes.

We expect to find other connections between graph generalizations and Data Mining based models.

We will continue ongoing work of Relational Structures, a generalization of graphs. Though our main motivation here is the Constraint Satisfaction Problem, from Artificial Intelligence, we have in mind that there are known connections with Data Mining algorithms.

1.3 Grammatical models

We aim to apply known grammatical concepts to sequence modelling, for instance categorial grammars.

1.4 Time series

We will apply techniques from automatic learning to the study of univariant and multivariant time series. The novelty here is that until now time series have been mainly studied with statistical techniques. We will include the following

- selection and tuning of the sequence learning models that are valid for time series
- inclusion of continuous and random variables (dynamical discretization)
- definition of new contrasts from the field of automatic learning.

1.5 Behaviour patterns in the study of subsequences

We want to complete the study of BPL, developed by Núñez in 2000 [38], that discovers temporal rules from the events and the static characteristics of an observed system and its surroundings. BPL first constructs a behaviour summary that is taken as training examples and then performs the construction of several behaviour trees that are used to predict the future.

Our particular objectives are to generalize the model to include base knowledge, to improve it so it detects chaotic behaviour, and to broaden the model with computed attributes.

1.6 Discovery of episodes

We want to find extensions and variants of existent (Data Mining) algorithms for episode search such as the following

- alternative algorithms for finding frequent sets
- combination of frequent sets with negative information

2 Level of success attained in the project

We list the objectives attained so far and the remaining tasks, organizing the information by the six families of models mentioned in the previous sections and finishing with data collection.

2.1 Compression mechanisms

We have completed the study of the equivalence, in the context of finite-state resource-bounds, of the compression ratio, the prediction log-loss performance and the effective dimension. This has been published in [19].

We have studied the case of worst case prediction of a sequence (instead of the usual best case performance), that we show it is equivalent to a well studied fractal dimension, packing dimension, equivalence that continues for all effectivizations and for space-bounded almost everywhere compression. This has been published as [1].

We have considered different prediction performance methods corresponding to dimension with “different scales” (scaled-dimension). This study has other applications in nonuniform complexity classes. This has been published in [27].

Finally, we have characterized the connection of this scaled-dimension with compressibility for the case of space-bounded computation. In this case compression can be formalized using Kolmogorov complexity. This has been published in [28].

We are currently working on the connection of polynomial-time prediction algorithms (and the corresponding effective dimension) with polynomial-time compression algorithms. We get an equivalence result by restricting compression algorithms to a kind that includes Lempel-Ziv. We expect to finish and write these results in the next months. We still don’t know if a full compression prediction equivalence holds in the context of time-bounds.

A student will shortly start a final project that will test the performance of Lempel-Ziv algorithms and some classical prediction schemes on a large amount of sequences. We are looking for a significant difference that will give hints on the improvement of some of the methods.

2.2 Generalizations of graphs

In the line of algorithms for construction and isomorphy of graphs we haven’t obtained substantial results. Related to this, Lozano et al. have studied the Maximum Common Embedded Subtree Problem in publication [30].

The objectives originally proposed for the TDIDT methods, formalization of the sampling algorithms and the use of voting, have been attained. The initial implementation of the methods has finished. We are still testing and comparing results. The method has been applied to the discovery of genes, this is included in publication [29].

The integration of decision trees with Markov hidden models in the MPSG model has successfully finished. We have an implementation that is still being tested.

In the subsection on Discovery of episodes we will mention several results where Galois connections and coproducts appear [15, 16]. This is a tiny step in our ambitious goal of connecting approaches.

Several publications have been obtained related to the Constraint Satisfaction Problem [12, 17, 18, 21].

2.3 Grammatical models

We have obtained results on the integration of syntax and semantics in the context of proof nets, contained in a publication on the geometry in grammar and circuits. This is publication [36].

We have developed a generalized logic of non concatenative operations on sequences. This is included on a paper on the discontinuity of type logical grammar, with applications for pronouns [35]. Related material has been presented in the Fields Workshop on Mathematical Linguistics in Ottawa, Canada.

Further results obtained in this models are related to proof nets, namely proof nets for the Lambek Calculus with brackets [23], non-associativity and balanced proof nets [22].

See also the results in publication [37].

2.4 Time series

Important advances have been achieved in the development of our new models for time series, that we have applied to climatology. See publications [34, 33].

2.5 Behaviour patterns in the study of subsequences

We have obtained a satisfactory generalization of BPL, presented in publications [42, 44, 43, 40, 41].

The model has been tested on atemporal noiseless problems (UCI Datasets) and with a classical problem, Stagger concepts, widely used when testing learning of concepts that change over time [24, 39].

2.6 Discovery of episodes

We summarize here the results obtained in the discovery of episodes, frequent sets and association rules.

We have proposed in [15] a new approach to the notions of closed subsequences and closed episodes, by the construction of new nonstandard Galois connections.

The extraction of unbounded episodes has been studied in [14].

The use of negative information in the search of frequent sets (initiated in FRESCO) has led to a new algorithm (see publication [25]).

We have adapted (strictly sequential) adaptive sampling to the computation of frequent sets, following a Best-First strategy and by the algorithm Ready-and-Go [5, 6].

The problem of Discrete Deterministic Data Mining, that is, inferring association rules with confidence rate 1 and support rate 0, has been analyzed from a different perspective. The obtained result coincides exactly with the Empirical Horn Approximation used in the area of Knowledge Compilation. This has been presented in publication [7].

We are working in learning sequences of Horn clauses, in the similar process of inferring data base dependencies and in the extraction of hybrid episodes of type sequence of sets. We have obtained first results in [9].

We have characterized multivalued dependencies and related expressions in a way that generalizes similar properties of functional dependencies and Horn clauses [8].

The use of graphs of concepts for the case of ordered data is explored in [16]. A software prototype of the algorithms in this paper and further extensions of it is currently under construction.

2.7 Other

The Final Project of Manuel Baena [3] dealt with the implementation of a data prospection model for the evaluation of work disability level. This work was done in collaboration with the Disability evaluation department in the Social Security National Institute. The same collaboration has lead to the PhD dissertation [13] and to the publication[4].

The work collected in publication [10] has been partially supported by this project and consists of definite versions of Support Vector Machine algorithms. The work was initiated within the FRESCO project.

Publications [2, 11, 20] have been partially funded by this project.

2.8 Data collection

The process of data collection for training and empirical validation of the model implementations is finished.

3 Performance indicators

3.1 Staff in training

There is a total of 9 PhD students that take part in the project, 5 of them are supported by the project (3 as research assistants, 2 as technical personnel).

- Barcelona: 3 PhD students (1 FPI research assistant).
- Málaga: 4 PhD students (1 FPI research assistant, 2 technicians).
- Zaragoza: 2 PhD students (1 research assistant).

A new senior member, Argimiro Arratia, is part of the Zaragoza subproject from August 2004.

3.2 Publications

The total number of publications in the project is 40 as of today (all of the references except four [26, 31, 32, 38]). These include:

- 5 international journal publications,
- 23 international conference publications,
- 4 chapters in specialized books,
- 2 national conference publications,
- 5 technical reports and unpublished manuscripts,
- 1 PhD dissertation.

3.3 Technological transfer and visibility

We have mentioned in the results section the existing collaboration with the Disability evaluation department in the Social Security National Institute. A software model currently in use has been implemented.

We are currently in preliminary contacts with three companies and institutions:

- The company Ravenpack, that works on analyzing a great amount of data through the Internet, is interested in applying the models developed within the MOISES project to the prediction of stock market events.
- The company Topdigital is interested in applying our models to the telecommunications domain.
- The Tourist Board of the Government of Andalucía is interested in Data Mining studies from tourism data such as opinion polls and complaint sheets. The Málaga subproject has elaborated a budget and is waiting to sign a collaboration contract.

Also, all our 25 conference contributions have been presented to a technical audience.

Glyn Morrill has been asked to write a contribution for the second edition of the Elsevier *Encyclopedia of Language and Linguistics*.

Argimiro Arratia will organize a special session on Logic and Computational Complexity in the Conference MAT.ES 2005 (Valencia, February 2005).

Elvira Mayordomo will organize a special session on Complexity in the Conference CiE 2005: New computational paradigms, in Amsterdam in June 2005.

We maintain a project public web page in
`http://webdiis.unizar.es/~elvira/moises.html`

3.4 Participation in international projects

E. Mayordomo is taking part in the NSF (USA government) grant "SGER: Multidisciplinary Aspects of Computation Theory", directed by J.H. Lutz.

Three researchers of this project are taking part in the initiative CIE (Computability in Europe) that aims to present a European Union proposal for a Research Training Network (RTNs). The proposal work will start from the Scientific conference "Computability in Europe 2005: New Computational Paradigms" that will be held in Amsterdam in June 2005.

3.5 Collaboration with other research groups

In the work on dimension and prediction algorithms, the contact with the group of Professor Lutz at Iowa State University is constant through frequent research visits.

The work around learning sequences of Horn clauses is done in cooperation with researchers from the University of Illinois at Urbana-Champaign and the Universidad del País Vasco.

Professor John M. Hitchcock from the University of Wyoming (USA) and professor Vinodchandran N. Variyam, from the University of Nebraska (USA) have visited Zaragoza to collaborate in this project.

References

[1] K. B. Athreya, J. M. Hitchcock, J. H. Lutz, and E. Mayordomo. Effective strong dimension in algorithmic information and computational complexity. In *Proceedings of the Twenty–First Symposium on Theoretical Aspects of Computer Science*, volume 2996 of *Lecture Notes in Computer Science*, pages 632–643. Springer-Verlag, 2004.

[2] A. Atserias and V. Dalmau. A combinatorial characterization of resolution width. In *Proceedings of the 18th IEEE Conference on Computational Complexity*, 2003.

[3] M. Baena. Prospección de datos: Estudios de diversas técnicas y su aplicación a datos sobre incapacidad. proyecto fin de carrera. Technical report, ETSIInformática, 2003.

[4] M. Baena, R. Morales, S. Cabuchola, and I. Santos. Prospección de datos sanitarios: Estudio de la incapacidad permanente. In *Inforsalud 2004*, pages 127–130, Madrid, 2004.

[5] J. Baixeries and G. Casas-Garriga. Sampling strategies for finding frequent sets. In *Actes des Journées Francophones d'Extraction et Gestion des Conaissances*, pages 159–170, 2003.

[6] J. Baixeries and G. Casas-Garriga. Sampling strategies for finding frequent sets. *Revue des sciences et technologies de l'information (RTSI) Serie RIA-ECA*, 17(1-2-3), 2003.

[7] J.L. Balcázar and J. Baixeries. Discrete deterministic data mining as knowledge compilation. In *Workshop on Discrete Mathematics and Data Mining in SIAM Int. Conf.*, 2003.

[8] J.L. Balcázar and J. Baixeries. Characterizations of multilevel dependencias and related expressions. In *Internacional Conference on Discovery Science*, 2004.

[9] J.L. Balcázar and G. Casas-Garriga. On horn axiomatizations for sequential data. In *Internacional Conference in Database Theory*, 2005. accepted.

[10] J.L. Balcázar, Y. Dai, and O. Watanabe. Provably fast training algorithms for support vector machines. Journal submission including the results from the three conference and workshop publications by the same authors.

[11] A. Bulatov, H. Chen, and V. Dalmau. Learnability of relatively quantified generalized formulas. In *Proceedings of the 15th International Conference on Algorithmic Learning Theory (ALT'04)*, 2004.

[12] A.A. Bulatov and V. Dalmau. Towards a dichotomy theorem for the counting constraint satisfaction problem. In *Proceedings of the 44th Symposium on Foundations of Computer Science (FOCS 2003)*, pages 562–573, 2003.

[13] S. Cabuchola. *Análisis de las patologías causantes de discapacidad laboral permanente mediante la aplicación de técnicas de inteligencia artificial y prospección de datos*. PhD thesis, Dept. de anatomía y medicina legal, Facultad de Medicina, Universidad de Málaga, 2003. Advisors Ignacio Miguel Santos Amaya and Rafael Morales Bueno.

[14] G. Casas-Garriga. Discovering unbounded episodes in sequential data. In *7th European Conference on Principles and Practice of Knowledge Discovery in Databases (PKDD'03)*, Cavtat-Dubvrovnik, Croatia, 2003.

[15] G. Casas-Garriga. Towards a formal framework for mining general patterns from structured data. In *Int. Workshop on Multirelational Datamining (MRDM 03) in KDD 03 Int. Conf.*, Washington, USA, 2003.

[16] G. Casas-Garriga and J.L. Balcázar. Coproduct transformations on lattices of closed partial orders. In *International Conference on Graph Transformation*, 2004.

[17] H. Chen and V. Dalmau. Looking algebraically at tractable quantified boolean formulas. In *Proceedings of the Seventh International Conference on Theory and Applications of Satisfiability Testing (SAT'04)*, Lecture Notes in Computer Science. Springer-Verlag, 2004.

[18] H. Chen and V. Dalmau. “smart” look-ahead arc consistency and the pursuit of csp tractability. In *Proceedings of the 10th International Conference on Principles and Practice of Constraint Programming (CP 2004)*, 2004.

[19] J. J. Dai, J. I. Lathrop, J. H. Lutz, and E. Mayordomo. Finite-state dimension. *Theoretical Computer Science*, 310:1–33, 2004.

[20] V. Dalmau and D. Ford. Generalized satisfiability with k occurrences per variable: a study through delta-matroid parity. In *Proceedings of the 28th International Symposium on Mathematical Foundations of Computer Science*, Lecture Notes in Computer Science. Springer-Verlag, 2003.

[21] V. Dalmau, A. Krokhin, and B. Larose. First-order definable retraction problems for posets and reflexive graphs. In *Proceedings of the 19th IEEE Symposium on Logic in Computer Science (LICS 2004)*, 2004.

[22] M. Fadda. Non-associativity and balanced proof nets. In *Proceedings of Categorial Grammars*, pages 46–58, Montpellier, 2004.

[23] M. Fadda and G. Morrill. The lambek calculus with brackets. In P. Scott, C. Casadio, and R. Seely, editors, *Language and Grammar: Studies in Mathematical Linguistics and Natural Language*. CSLI, 2004. In press.

[24] R. Fidalgo. Herramienta para la construcción de árboles de decisión. proyecto fin de carrera. Technical report, ETSIInformática, 2003.

[25] I. Fortes, J.L. Balcázar, and R. Morales. Bounding negative information in frequent sets algorithms. In *4th International Conference DS 2001*, Lecture Notes in Artificial Intelligence, pages 50–58, 2001.

[26] J. M. Hitchcock. Fractal dimension and logarithmic loss unpredictability. *Theoretical Computer Science*, 304(1–3):431–441, 2003.

[27] J. M. Hitchcock, J. H. Lutz, and E. Mayordomo. Scaled dimension and nonuniform complexity. *Journal of Computer and System Sciences*, 69:97–122, 2004.

[28] J.M. Hitchcock, M. López-Valdés, and E. Mayordomo. Scaled dimension and the Kolmogorov complexity of Turing-hard sets. In *Proceedings of the 29th International Symposium on Mathematical Foundations of Computer Science*, volume 3153 of *Lecture Notes in Computer Science*, pages 476–487. Springer-Verlag, 2004.

[29] J. Jerez, J.A. Gómez-Ruiz, G. Ramos, J. Múñoz, and E. Alba-Conejo. A combined neural network and decision trees model for prognosis of breast cancer relapse. *Artificial Intelligence in Medicine*, 27:45–63, 2003.

[30] A. Lozano and G. Valiente. On the maximum common embedded subtree problem for ordered trees. In C. Iliopoulos and T. Lecroq, editors, *String Algorithmics*, pages 155–169. King's College London Publications, 2004. In press.

[31] J. H. Lutz. The dimensions of individual strings and sequences. *Information and Computation*, 187:49–79, 2003.

[32] E. Mayordomo. A Kolmogorov complexity characterization of constructive Hausdorff dimension. *Information Processing Letters*, 84(1):1–3, 2002.

[33] L. Mora, R. Ruiz, and R. Morales. Modelo para la selección automática de componentes significativas en el análisis de series temporales. In *Caepia'03*, 2003.

[34] L. Mora-Lopez, R. Morales, M. Sidrach de Cardona, and F. Triguero. Probabilistic finite automata and randomness in nature: a new approach in the modelling and prediction of climatic parameters. In *International Environmental Modelling and Software Congress*, pages 78–83, Lugano, Suiza, 2002.

[35] G. Morrill. On anaphora in type logical grammar. In G.-J. Kruiff and R.T. Oehrle, editors, *Resource Sensitivity and Binding*. Kluwer Academic Publishers, Dordrecht. To appear.

[36] G. Morrill. Geometry of language and linguistic circuitry. In P. Scott, C. Casadio, and R. Seely, editors, *Language and Grammar: Studies in Mathematical Linguistics and Natural Language*. CSLI, 2004. In press.

[37] G. Morrill and A. Gavarró. On aphasic comprehension and working memory load. In *Proceedings of Categorial Grammars*, pages 259–287, Montpellier, 2004.

[38] M. Núñez. Learning patterns of behaviour by observing system events. In *Proceedings of the 11th European Conference on Machine Learning*, volume 1810 of *Lecture Notes in Computer Science*, pages 323–330. Springer-Verlag, 2000.

[39] M. Núñez and R. Fidalgo. Aprendizaje de árboles de decisión temporales a partir de eventos. Technical Report ITI-2003-01, Departamento de Lenguajes y Ciencias de la Computación, Universidad de Málaga, 2003.

[40] M. Núñez, R. Fidalgo, and R. Morales. Discovering temporal patterns from events and other multivariate data. In *Proceedings of the Congress Euro Electromagnetics (EUROEM 2004)*, Magdeburg (Germany), 2004.

[41] M. Núñez, R. Fidalgo, and R. Morales. Reducing potential risks by preventing events: A case study. In *Proceedings of the IFAC Congress on Management and Control of Production and Logistics (MCPL-2004)*, Oxford, UK, 2004.

[42] M. Núñez and R. Morales. Learning prediction knowledge for nonlinear systems. Technical Report ITI-02-4, Departamento de Lenguajes y Ciencias de la Computación, Universidad de Málaga, 2002.

[43] M. Núñez and R. Morales. Incorporating prediction facilities to autonomous systems. In *Autonomous Agents & Multi Agent Systems*. IEEE Computer Society Press, 2004.

[44] M. Núñez, R. Morales, and F. Triguero. Automatic discovery of rules for predicting network management events. *IEEE Journal on Selected Areas in Communications*, 20(4), 2002. Special Issue: Recent Advances in Fundamentals of Network Management.