MŰEGYETEM 1782

Budapest University of Technology and Economics
Faculty of Electrical Engineering and Informatics
Department of Control Engineering and Information Technology

MODELING CLIENT INFORMATION SYSTEMS USING BAYESIAN NETWORKS

Main results of the PhD Thesis

Gábor Vámos

Advisor: Dr. Ákos Nagy PhD.
Department of Control Engineering and Information Technology

Budapest 2006

Introduction

The utility companies, the financial service providers, and the telecommunication corporations maintain their complex customer information systems (CIS) primarily for periodical invoicing purposes. From the aspect of the PhD dissertation, a general CIS consists of two main components: one of them is the database stored on technical instruments and the other one is the team of employees managing the customer data. The data knowledge, which includes the correlations between the data, provides an objective description of the physical world. The individuals hold possession of practical and/or theoretical knowledge about the business processes in from of a subjective explanation about the problem domain. Based on these two knowledge types an expert system can be developed to support the complex decisions of the owner institute of the CIS.

The intention of the CIS model generation process is to find an abstract representation of the information coded in the database and that of the human knowledge accumulated as experience in the minds of employees. Since both of the relevant knowledge types are uncertain and redundant, the mathematical background we use to construct knowledge based expert systems is provided by the probability theory.

Probabilistic models of real world phenomena are increasingly used in commercial applications, especially in decision support systems and in different types of fault or fraud detection tasks. Such applications have two essential, usually independently considered components: the knowledge base (abstract description of real world phenomena, i.e. a model) and the model evaluation engine [Russel03].

In the case of an uncertain problem domain, the model is often given as a joint probability distribution function (PDF) over a finite set of variables. For discrete variables, the joint probability function may be stored in a table. This may result simple inference engines at the price of huge required storing capacity which grows exponentially with the number of variables and their possible values. In fact, the required storage capacity impedes the direct use of joint probability tables in real applications.

Bayesian Networks offer an efficient representation structure using conditional probability tables obtained from the factorization of the joint probability function. This factorization is implied by a set of independence relations which is represented by a directed acyclic graph (DAG). This graphical structure helps to integrate the two vital knowledge sources when creating the probabilistic model.

Further major advantages of Bayesian Networks are their reusability and application independent availability. The reusability means that a model version, described with a DAG and the numerical components, can be modified according to the alteration of the problem domain. The application independency expresses that task-oriented manner appears solely in the model queries process implemented by probabilistic inferences. Namely the choice of the input and output fields – in the implemented knowledge model – are not fixed to a specific inference situation.

The dissertation is organized as follows. The first part gives a general survey about the probabilistic modeling techniques focusing on the Bayesian Networks. These sections deal with the properties of the Bayesian Network models and give a review about the most relevant mathematical algorithms (probabilistic inferences and learning). The second part introduces the novel scientific results of the research. The primary aim of the working-out was to support directly the knowledge engineers in their workflows.

The aim of the dissertation

Widespread utilization of Bayesian Networks in practical applications is restricted by the enormous demand of computational resources. It has been proven that the probabilistic inference process [Chickering03] and the learning task [Cooper90][Dagum93] in Bayesian Networks with general DAG structure are NP-Hard. The required computing capacity resulted by the modeling of large problem domains can be kept under control by the use of heuristic algorithms in order to create a suitable balance between the model resolution and the computing cost [Bouckaert95][Kjaerulff90].

The main objective of the proposed model development algorithms:

- acquisition and integration of multiple expert knowledge weighting the individual competence properties and the data knowledge;
- determination of the structural model component which decisively affects the computational cost of the model;
- optimization of the computational requirements enhancing the relationship between the model generation and the inference processes.

Summary of new scientific results

Thesis 1: I have developed a new methodology for the probabilistic modeling process based on the integration of data knowledge and expert knowledge [2][12].

A knowledge based model is headed throughout the development task by the Knowledge Engineering Process (KEP). Many KEP methodology possibilities are available to implement a model [Taboada03]. Most of them are focusing especially on the deterministic description of the expert knowledge without redundancy.

However, in the case of the mapping of a CIS into a model, multiple uncertain knowledge sources are necessary to be applied jointly to observe the behavior of real world objects. It follows that the published deterministic KEP methods are not suitable for the representation of a CIS model in this situation. Scientific papers published in the last 25 years in the field of probabilistic models are engaged mostly in several algorithm-specific problems and deal only occasionally with comprehensive methodology issues.

Thesis 1 specifies an interdisciplinary methodology for the problem of probabilistic modeling with Bayesian Networks. The required computing capacity of model development and inference tasks can be reduced on the basis of this methodology.

Standard UML tools were used for the specification of this methodology. The significant elements:

- categorization of problem domains based upon their level of observability;
- classification of human roles and technical entities in the modeling process;
- functional definition of sub-processes;
- hierarchical description of sub-processes with detailed specification of the interfaces between them;
- definition of states and state-transitions of the modeling process giving the facilities of backtracking;
- description of elemental increments in the graphical model component (GUI).

> Thesis 2: I have developed a new incremental algorithm to generate the structural model component of a Bayesian Network using multiple knowledge sources [5].

Based solely on the data knowledge, the model development process of a CIS may lead to an NP-hard problem because the search space cardinality (the number of potential Bayesian Network structures) is growing super-exponentially with the number of random variables [Robinson76].

Generally, the human knowledge comes from different experts. Therefore the description of the problem domain is redundant due to the diverse interpretation abilities of individual professionals [Feigenbaum84]. The redundancy and uncertainty of the expert knowledge and the NP-hard nature of the data learning process imply that both knowledge sources (data and expert knowledge together) are used in the development of the knowledge based models.

The proposed algorithm aims to keep in control the computing costs of the structural learning and to give a solution to handle the uncertainty in the expert knowledge exploration task.

Significant properties:

- the algorithm decomposes the modeling process into iteration steps: the structural model component is developed in every iterations by the extraction and representation of expert knowledge;
- at the end of each iteration steps the generated structures are rated objectively using the data knowledge;
- the objective evaluations are performed based on the AIC - Akaike Information Criterion – learning metric [Akaike74], wherein the calculation of fitting and complexity are detached sharply [Friedman91], therefore the balance between the model accuracy and the complexity of its evaluation can be easily tuned;
- the elicitation of more experts' knowledge is executed in run-time: it has been proven by case-studies that the development-time integration of different knowledge sources effects inconsistencies in the independency relations [Richardson03][Williamson01];
- the expert knowledge is considered by the algorithm in global aspects.

Thesis 3: I have worked out a new algorithm to generate the structural model component of a Bayesian Network model based on the visualization of relevant new model development states [1].

The elicitation and integration of expert knowledge into the structural model component can be mapped directly by the human expert or indirectly by the involvement of the knowledge engineer. In both cases, the practical approach of modifying the DAG structure is based on a graphical visualization instead of the direct alteration of the mathematical representation (adjacency matrix, adjacency list etc.)

There are several commercial and research tools designed for Bayesian Network model authoring and testing. A major functional deficiency of these tools is their incapacity to display in advance any information about the effect on DAG property of an initiated edge increments. The DAG property is checked - using any graph discovery algorithm - just after the specification of the vertex pairs belonging to a new edge insertion operation. The model development is carried out with random attempts to insert new edges, because the possible forward steps in the model development states are not visualized. It presents difficulties especially in the case of large graphs with high densities hence the efficiency of collaboration between the knowledge engineer and experts is lowered by the frequent failures of edge insertion attempts.

The proposed algorithm aims to improve the efficiency of the DAG development process in Bayesian Network models. It visualizes the possible edge pairs (that do not generate directed circles in the graph) in the midst of edge inserting operation. In this way, the relevant model development alternatives become unambiguously transparent.

Major properties:

- the edge inserting operation is supported by a special structure (reachable_from matrix) which is maintained to represent the reachablity of each node from the other nodes;
- the thesis gives an upper limit for the size of reachable_from matrix and defines an algorithm to generate this matrix;
- it was shown that the required computing capacity of matrix generation does not embarrass the interactions (between the knowledge engineer and the experts) in the graphical model development.

Thesis 4: I have worked out a new algorithm to refine the structural model component of a Bayesian Network model based on edge removal from the network [6][11].

According to the regular approach, the result of the model generation process is a standalone structure (DAG and the conditional probability tables) which is independent from the probabilistic inference, so the created model perfectly supports the portability among applications. Most often, the type of inference method is known and fixed. Under such circumstances the general applicability is not required for the model, but the balance between the model accuracy and the required computing capacity of its evaluation is important for the sake of efficient functioning of model.

In the case of the edge removal operation the fixation of inference algorithm is particularly expedient. The edge removal operations are considered characteristically at the last phases of the incremental model development process when the exploration of object relationships in the model space turns out to be complete. By this time, some early edges can represent statistically weak relationships compared to some newly inserted edges. If the network complexity converges to its maximal level, then some old weak edges have to be removed before a new strong edge is inserted into the network. The knowledge engineer is responsible for the trade-off between the model resolution and the complexity of its inference. This balance should be more easily found if the knowledge engineer could receive an immediate feedback about the computational resources required to evaluate his or her actual probabilistic model.

Significant properties:

- this thesis makes possible to establish a trade-off between the criteria of the development and the application phases of probabilistic models represented by Bayesian Networks;
- the proposed algorithm is based on the widely utilized AIC learning metric [Friedman91] and on the PPTC inference method [Gaag97];
- the rating function in the new algorithm uses the clique size generated by the triangular heuristics, so the changes in the model complexity occurred as the effect of the actual model increment are evaluated with the prospective changes in the required computing capacity of inference algorithm.

Practical applications

The first and second theses were worked out and applied in the project framework of „IKTA4-042 Detecting the statistical structure of large databases and development of a human interface for the generated knowledge base" leaded by the BME Department of Control Engineering and Information Technology and by the DSS Consulting Kft. [11][12][13][14][15].

The algorithms of the third and fourth theses were presented in a prototype software designed using MATLAB R14 tool [6]. The implemented graphical software interface is able to support the visualization and the modification of DAG structures and the numerical data of conditional probability tables. The source code contains 4000 lines including the function libraries and comments.

According to the methodology of this dissertation, several customer information systems were mapped into probabilistic models during the period of 2002-2007 [2][5][8][17][16][18]. The main goals of these applications were to model fraudulent consumer behavior among clients of utility companies with the ambition to enlarge the hit rate among controlled consumers and to decrease the unit cost of a successful detection of frauds. I would like to express my sincere gratitude to our industrial partners for their confidence and valuable support: ÉMÁSZ Nyrt., ELMŰ Nyrt., E.ON Zrt.

Applications in education

Real-time systems and networks (VIFO4364)
Software Engineering, Semester 8, 5 Credits, 4 contact hrs/week, the topic covers 50 percent of the subject
URL: https://www.vik.bme.hu/kepzes/targyak/VIFO4364/

Integrated Planning and Control of Energetic Systems (VIIIM409)
Software Engineering, Semester 8, 5 Credits, 4 contact hrs/week, the topic covers 50 percent of the subject
URL: https://www.vik.bme.hu/kepzes/targyak/VIIIM409/

Enterprise and Manufacturing Control Systems Laboratory
Software Engineering, Semester 9, 3 Credits, 2 contact hrs/week, the topic covers 33 percent of the subject
URL: https://www.vik.bme.hu/kepzes/targyak/VIFO5305/

References

[Russel03] S. J. Russel, P. Norvig: *Artificial Intelligence: A Modern Approach*, Prentice Hall Inc., 2nd edition, 2003.

[Chickering03] D. M. Chickering, C. Meek and D. Heckerman: Large-Sample Learning of Bayesian Networks is NP-Hard, *In Proceedings of Nineteenth Conference on Uncertainty in Artificial Intelligence*, *pp 124-133*, Morgan Kaufmann, Acapulco, Mexico, 2003.

[Cooper90] G. F. Cooper: The Computational Complexity of Probabilistic Inference Using Bayesian Belief Networks, *Artificial Intelligence, 42 (2-3), pp. 393--405*, 1990.

[Dagum93] P. Dagum, M. Luby: Approximate Probabilistic Inference in Bayesian Networks is NP Hard, *Artificial Intelligence, 60, pp. 141-153*, 1993.

[Bouckaert95] R. R. Bouckaert: Bayesien Belief Networks: From Construction to Inference, *PhD Thesis, University to Ultrecht*, 1995.

[Kjaerulff90] U. Kjaerulff: Triangulation of Graphs - Algorithms Giving Small Total State Space, *Technical Report TR R 90-09, Department of Mathematics and Computer Science*, Strandvejen, Aalborg, Denmark, 1990.

[Taboada03] M. Taboada, M. Argüello, J. Des, J. Mira: Building Knowledge-Based Intelligent Systems by Reusing, *In Innovations in KE, International Series on Advanced Intelligence, Vol. 82, pp. 1-30*, 2003.

[Robinson76] R. D. Robinson: Counting Unlabeled Acyclic Digraphs, *In Proceedings Australian Conference on Combinatorial Mathematics*, *Vol. 5, pp. 28-43*, 1976.

[Feigenbaum84] E. A. Feigenbaum: Knowledge Engineering: The Applied Side of Artificial Intelligence, *Annals of the New York Academy of Sciences, 426, pp. 91–107*, 1984.

[Akaike74] H. Akaike: A new Look at the Statistical Model Identication, *IEEE Transactions on Automatic Control 19, pp. 716-722*, 1974.

[Friedman91] J. Friedman: Multivariate adaptive regression splines (with discussion), *Annals of Statistics 19, pp.* 1-67, 1991.

[Richardson03] M. Richardson, P. Domingos: Learning with Knowledge from Multiple Experts, *In Proceedings of the Twentieth International Conference on Machine Learning (ICML-2003), pp. 624-631*, Washington DC, 2003.

[Williamson01] J. Williamson: Foundations for Bayesian networks, *In eds. David Corfield & Jon Williamson Foundations of Bayesianism, pp. 75–115*, *Kluwer Applied Logic Series*, 2001.

[Gaag97] L. C. van der Gaag, H. L. Bodlaender: Comparing Loop Cutsets and Clique Trees in Probabilistic Inference, *Technical Report UU-CS-1997-42, Utrecht University: Information and Computing Sciences*, 1997.

Publications related to the theses

Journal papers published in foreign language

[1] Vámos: A Matlab Graphical Tool to Support Knowledge Engineering, *Production Systems and Information Engineering, Vol. 4, pp. 81-93*, Miskolc University Press, 2006, ISSN 1785-1270

Journal papers published in Hungarian

[2] Vámos-Beck: Fogyasztók viselkedésének tudásalapú modellezése, *Elektrotechnika 99. évf. 2006/3, pp. 17-18*, 2006, HUISSN 0367-0708

Conference papers published in foreign language

[3] Kovács-Kiss-Nagy-Vámos: Early Detection System for Vegetation Fire in the Aggtelek National Park, *Proceedeings of IEEE TEHOSS'2005, pp. 493-498*, Gdańsk, 2005.09.28-30. ISBN 83-917681-9-8.

[4] Kovács-Kiss-Nagy-Vámos: Autonomous Fire Detection and Warning System for Early Recognition of Vegetation Fires, *Proceedings of MicroCad 2005 International Scientific Conference, pp. 241-246*, Miskolc, 2005.03.10-11. ISBN 963-661-646-9.

[5] Arató-Kiss-Vajta-Vámos: Fraudulent Consumer Behaviour Analysis and Detection for Utility Companies, *Proceedeings of 2nd Conference on Information Technology, pp. 37-43*, Gdańsk, 2004.05.16-18. ISBN 83-917681-5-5.

[6] Vámos-Nagy-Kiss: GUI Tool for Manipulate and Refine Bayesian Networks, *Proceedings of Second Hungarian Conference on Computer Graphics and Geometrics, pp. 65-70*, Budapest, 2003.06.30.-2003.07.01.ISBN 963-420-766-9.

[7] Vámos-Nagy-Kiss: Bayesian Network Based Modeling for Flaw Detection in Metallic Fusion Welds USING X-Ray Images, *Proceedings of 4th Workshop on European Scientific and Industrial Collaboration Promoting Advanced Technologies in Manufacturing, pp. 181-186*, Wesic, Miskolc, 2003.05.28-30. ISBN 963-961-570-5.

[8] Vámos-Nagy-Kiss: Creating Bayesian network based probabilistic models for practical applications, *Proceedings of MicroCad 2003 International Scientific Conference, pp. 163-168*, Miskolc, 2003.03.6-7. ISBN 963-661-547-0.

[9] Vámos-Loványi-Nagy-Kiss: Flaw Detection in Metallic Fusion Welds on X-ray Images Using Bayesian Networks, *Proceedings of First Hungarian Conference on Computer Graphics and Geometrics, pp. 118-123*, Budapest , 2002.05.28-29. ISBN 963-420-718-9.

[10] Vámos-Nagy: Cardholder Identification Systems based on Behaviour Analysis and Biometric Methods, *Proceedings of MicroCad 2001 International Scientific Conference, pp. 177-182*, Miskolc, 2001.03.01-02. ISBN 963-661-457-1.

Technical Reports

[11] Vámos-Kiss: Nagy adatbázisok valószínűségi struktúrájának feltárása és humán interfész kifejlesztése – Részletes Kutatási Jelentés
IKTA4-042, 2002.12.15

[12] Vámos-Kiss-Aszalós: Nagy adatbázisok valószínűségi struktúrájának feltárása és humán interfész kifejlesztése - Módszertani Tanulmány
IKTA4-042, 2002.07.1.

Presentations without published paper

[13] Vámos-Nagy: Nagy adatbázisok valószínűségi struktúrájának feltárása és humán interfész kifejlesztése – Projektzáró prezentáció, IKTA4 FÓRUM, Budapest, 2004.04.15

[14] Kiss-Nagy-Egri-Vámos: Nagy adatbázisok valószínűségi struktúrájának feltárása és humán interfész kifejlesztése – Projektbemutató prezentáció, IKTA4 FÓRUM, Budapest, 2002.11.20.

[15] Aszalós-Vámos: Nagy adatbázisok valószínűségi struktúrájának feltárása és humán interfész kifejlesztése – Projektbemutató prezentáció, IKTA4 FÓRUM, Budapest, 2002.02.07.

Research reports

[16] Vámos-Kiss-Vajta: A lakossági ügyfélkör ellenőrzési címlistáinak generálása, Projektzáró dokumentum, 2004.11.25.

[17] Vámos-Kiss-Vajta: Az áramszámlázási adatbázis hihetőség-vizsgálata, Projektzáró dokumentum, 2003.09.25.

Conference paper published as an abstract only

[18] Vámos-Beck: Szabálytalan vételezés felderítésének támogatása az ügyfelek viselkedésének valószínűségi modellezésével, *Magyar Elektrotechnikai Egyesület 52. Jubileumi Vándorgyűlés Konferencia kiadványa*, 2005