BMAW 2016

The Thirteenth UAI Bayesian Modeling Applications Workshop

June 25, 2016

New York City, USA

Workshop Proceedings

Rommel Novaes Carvalho and Kathryn Blackmond Laskey, Eds.

http://ceur-ws.org/Vol-1663/

Preface

The Bayesian Modeling Applications Workshop (BMAW) has been held in conjunction with the annual Conference on Uncertainty in Artificial Intelligence thirteen times since 2003. The workshop brings together researchers and practitioners who apply the technologies pioneered by the UAI community to address important real-world problems in a diverse set of fields. The workshop fosters discussion on the challenges of building applications, such as understanding and addressing stakeholder needs; integrating Bayesian models and tools into larger applications; validating models; interacting with users; construction of models through knowledge elicitation and learning; agile model and system development strategies; and deploying and managing Web based Bayesian applications.

The theme of the Workshop has adapted from year to year, as real-world problems change and technologies evolve to meet them. The frenzy to apply conventional machine learning methods for commercial applications has the danger of overwhelming Bayesian methods where they might be best applied. Bayesian methods face a similar challenge to the one they faced a decade ago by this community: To demonstrate their timeliness in the current environment of intelligent systems and a long tail of related decision and prediction tasks. This Workshop demonstrates that through several tools and current applications of Bayesian methods.

A call for papers encouraged submissions in a variety of domains, but not limited to any specific vertical market or discipline. Submissions were expected to foster discussion of critical issues within the community of practice. There were 9 submissions. Each submission was reviewed by at least three program committee members. Eight papers were accepted and presented at the Workshop. Seven of these appear full length in these proceedings. One appears as extended abstract to facilitate future publication. In addition, three invited speakers have blessed the Workshop with the presentation of their poster paper accepted at the main conference.

The Thirteenth Annual BMAW was held on June 25, 2016, in New York City, NY, USA. About 30 people attended the Workshop, which consisted of eleven paper presentations, questions, and the accompanying discussions. Papers and presentations addressed Bayesian learning algorithms, tools, and several applications involving medical, government, tax, robotics, soccer, corruption, and education domains. We are grateful to the paper authors and presenters for their contributions, and to the program committee members for their careful efforts reviewing and commenting on submissions. We also appreciate the help EasyChair has always provided us and the organizational support provided by the UAI conference organizers, without whom the workshop would not be possible. Finally, we also thank the authors of the main conference for accepting our invitation to present an invited talk.

June 2016
New York City, NY, USA

Rommel Novaes Carvalho
Kathryn Blackmond Laskey
Workshop Co-Chairs

Table of Contents

Full Papers

Extended Abstract

Abstracts of Invited Talks from UAI 2016

Program Committee

Russell Almond	Florida State University
Rommel Carvalho	University of Brasília / Brazil's Office of the Comptroller General
Feng Chen	SUNY Albany
Paulo Costa	George Mason University
Pablo González	Instituto de Investigaciones Electricas Mexico
Sajjad Haider	Institute of Business Administration
Arjen Hommersom	Open University of the Netherlands
Oscar Kipersztok	The Boeing Company
Helge Langseth	Norwegian University of Science and Technology
Kathryn Laskey	George Mason University
Ole Mengshoel	Carnegie Mellon University
Tomas Singliar	Microsoft
V Anne Smith	University of St Andrews
Luis Enrique Sucar	INAOE

A Risk Calculator for the Pulmonary Arterial Hypertension Based on a Bayesian Network

Jidapa Kraisangka & Marek J. Druzdzel *
Decision System Laboratory,
School of Information Sciences,
University of Pittsburgh,
Pittsburgh, PA

Raymond L. Benza
Advanced Heart Failure, Transplant,
MCS and Pulmonary Hypertension
Allegheny Health Network
Allegheny General Hospital
Pittsburgh, PA

Abstract

Pulmonary arterial hypertension (PAH) is a severe and often deadly disease, originating from an increase in pulmonary vascular resistance. Its prevention and treatment are of vital importance to public health. A group of medical researchers proposed a calculator for estimating the risk of dying from PAH, available for a variety of computing platforms and widely used by health-care professionals. The PAH Risk Calculator is based on the Cox's Proportional Hazard (CPH) Model, a popular statistical technique used in risk estimation and survival analysis, based on data from a thoroughly collected and maintained Registry to Evaluate Early and Long-term Pulmonary Arterial Hypertension Disease Management (REVEAL Registry). In this paper, we propose an alternative approach to calculating the risk of PAH that is based on a Bayesian network (BN) model. Our first step has been to create a BN model that mimics the CPH model at the foundation of the current PAH Risk Calculator. The BN-based calculator reproduces the results of the current PAH Risk Calculator exactly. Because Bayesian networks do not require the somewhat restrictive assumptions of the CPH model and can readily combine data with expert knowledge, we expect that our approach will lead to an improvement over the current calculator. We plan to (1) learn the parameters of the BN model from the data captured in the REVEAL Registry, and (2) enhance the resulting BN model with medical expert knowledge. We have been collaborating closely on both tasks with the authors of the original PAH Risk Calculator.

1 Introduction

Pulmonary arterial hypertension (PAH) is a fatal, chronic, and life-changing disease originating from an increase in pulmonary vascular resistance, and leading to high blood pressure in the lung (Benza et al., 2010; Subias et al., 2010). Patients with PAH suffer from shortness of breath, chest pain, dizziness, fatigue, and possibly other symptoms depending on the progression of disease (Hayes, 2013). Currently, there is no cure for PAH and treatment is often determined based on the symptoms. With an early diagnosis and proper treatment, patients' lives can be extended by five or more years.

With the long-term goal to characterize the clinical course, treatment, and predictors of outcomes in patients with PAH in the United States, a group of medical researchers established a Registry to Evaluate Early and Long-term Pulmonary Arterial Hypertension Disease Management (REVEAL Registry) (Benza et al., 2010). The REVEAL registry is quite likely the most comprehensive collection of data of patients suffering from PAH and it has led to interesting insights improving the diagnosis, prediction, and treatment of PAH. One of the prominent applications of the REVEAL Registry is the PAH Risk Calculator (Benza et al., 2012), a statistical model learned from the REVEAL Registry data and predicting the survival of patients at risk for PAH. A computer implementation of the PAH Risk Calculator is available for a variety of computing platforms and widely used by health-care professionals (see http://www.pah-app.com/ for more information).

The PAH Risk Calculator is based on the Cox's Proportional Hazard (CPH) model (Cox, 1972), a popular statistical technique used in risk estimation and survival analysis. One weakness of this approach is that the underlying model can be only learned from data and is not readily amenable to refinement based on expert knowledge. Another possible weakness is that the CPH model rests on several assumptions simplifying the interactions between the risk factors and the disease. While these assumptions are reasonable and the CPH model has been successfully used for decades,

* Also Faculty of Computer Science, Bialystok University of Technology, Bialystok, Poland

it is interesting to question them with a possible benefit in terms of model accuracy.

In this paper, we propose an alternative approach to calculating the risk of PAH that is based on a Bayesian network (BN) (Pearl, 1988) model. BNs are acyclic directed graphs in which vertices represent random variables and directed edges between pairs of vertices capture direct influences between the variables represented by the vertices. A BN captures the joint probability distribution among a set of variables both intuitively and efficiently, modeling explicitly independences among them. A representation of the joint probability distribution allows for calculation of probability distributions that are conditional on a subset of variables. This typically amounts to calculating the probability distributions over variables of interest given observations of other variables (e.g., probability of one-year survival given a set of observed risk factors). There is a well developed theory expressing the relationship between causality and probability and often the structure of a BN is given a causal interpretation. This is utmost convenient in terms of user interfaces, notably knowledge acquisition and explanation of results. The first step in our work has been to create a BN model that mimics the CPH model at the foundation of the current PAH Risk Calculator. In this, we use the BN interpretation of the CPH model proposed by Kraisangka and Druzdzel (2014). Our BN-based calculator reproduces the results of the current PAH Risk Calculator exactly.

Because Bayesian networks do not require the assumptions of the CPH model and can readily combine data with expert knowledge, we expect that our approach will eventually lead to an improvement over the current PAH Risk Calculator. Our mid- to long terms plans include (1) learning the parameters of the BN model directly from the data captured in the REVEAL Registry, and (2) enhancing the resulting BN model with medical expert knowledge. We are collaborating on both tasks with the team maintaining the REVEAL Registry and the authors of the original PAH Risk Calculator.

The remainder of this paper is structured as follows. Section 2 describes the problem of PAH, the CPH model, and the PAH Risk Calculator. Sections 3 and 4 describe application of Bayesian networks to risk estimation and the proposed BN-based PAH Risk Calculator. Finally, Section 5 describes our conclusions and future work.

2 Pulmonary Arterial Hypertension

This section introduces some facts related to the pulmonary arterial hypertension (PAH), notably its risk factors, the Cox's Proportional Hazard (CPH) model, and the PAH Risk Calculator based on the CPH model.

PAH Risk Factors

Risk can be defined as the rate of an occurrence of a particular disease or adverse event (Irvine, 2004). Although PAH can occur at any age, in any races, and any ethnic background (Hayes, 2013), there are risk factors that make some people more susceptible. For example, females are at least two and a half times more susceptible than men to idiopathic PAH. Recently, medical care professionals treating PAH have relied on existing patient registries to understand PAH better. Several risk factors have been identified and used to develop prognostic models for guiding their therapeutic decision making. For example, a study based on the Registry to Evaluate Early and Long-Term Pulmonary Arterial Hypertension Disease Management (REVEAL) (Benza et al., 2010) extracted several demographic, functional, laboratory, and hemodynamic parameters associated with patient survival in PAH (Benza et al., 2012) by means of a multivariate Cox's proportional hazard model (CPH) (discussed in more detail in the following section). By developing a prognosis model, physician can access a short-term and long-term patient survival in the context of current treatment and clinical variables (Benza et al., 2012). Although prognostic tools for patient survival have improved the quality of predictions, the models are still imperfect and more research is needed on improving them.

Cox's Proportional Hazard Model

Hazard is a measure of *risk* at a small time interval t, which can be considered as a rate (Allison, 2010). In survival analysis, the hazard function can be represented by probability distributions (e.g., exponential distribution) or can be modeled by regression techniques. The Cox's proportional hazard model (CPH) (Cox, 1972) is a set of regression methods used in the assessment of survival based on its risk factors or explanatory variables. The probability of an individual surviving beyond time t can be estimated with respect to a hazard function (Allison, 2010). As defined originally by Cox (1972), the hazard regression model is expressed as

$$\lambda\left(t\right) = \lambda_0\left(t\right) \exp^{\beta' \cdot \mathbf{X}} \ . \tag{1}$$

This hazard model is composed of two main parts: the baseline hazard function, $\lambda_0\left(t\right)$, and the set of effect parameters, $\beta' \cdot \mathbf{X} = \beta_1 X_1 + \beta_2 X_2 + ... + \beta_n X_n$. The baseline hazard function determines the risks at an underlying level of explanatory variables, i.e., when all explanatory variables are absent. The βs are the coefficients corresponding to the risk factors, $\mathbf{X}$. According to Cox (1972), this $\lambda_0\left(t\right)$ can be unspecified or can follow any distribution and be estimated from data.

The application of the CPH model relies on the assumption that the hazard ratio of two observations is constant over time (Cox, 1972). For example, a hazard ratio of a group of

PAH patients having renal insufficiency to a group of PAH without renal insufficiency (control/baseline group) is estimated as 1.90. This assumption means that patients with renal insufficiency always have a 90% higher risk for dying from PAH than patients without renal insufficiency by Cox's assumptions. The ratio of two hazards is defined as γ:

$$\gamma = \frac{\lambda_2(t)}{\lambda_1(t)} = \frac{\exp(\beta' X_2)}{\exp(\beta' X_1)} \ . \quad (2)$$

If the risk factors X are binary, their value could be expressed as *presence* $(X = 1)$ or as *absence* or *baseline* $(X = 0)$ of the risk factor. Once, we know the hazard ratio of one group toward another group, we can estimate the survival probability (Casea et al., 2002) by

$$S(t) = S_0(t)^{\gamma} \ . \quad (3)$$

$S_0(t)$ is the baseline survival probability estimated from the data, i.e., when all risk factor are absent or at their baseline value $(X = 0)$ at any time t, while γ is hazard ratio of an interested group to the baseline group. In other words, the survival probability of any patients relative to the baseline group can be estimated from

$$S(t) = S_0(t)^{\exp^{\beta' \cdot \mathbf{x}}} \ . \quad (4)$$

An example of CPH model used as a prognosis model for PAH patients is from the REVEAL Registry Risk Score Calculator (Benza et al., 2012). The model, including 19 risk factors, was developed to predict a one-year survival probability. The main survivor function is

$$S(t=1) = S_0(1)^{\exp^{\beta' \cdot \mathbf{x}_\gamma}} \ , \quad (5)$$

where $S_0(1)$ is the baseline survivor function of 1 year (0.9698) and γ in this equation is the shrinkage coefficient after model calibration (0.939) (Benza et al., 2010). The risk factors X (listed in Table 1) included PAH associated with portal hypertension (APAH-PoPH), PAH associated with connective tissue disease (APAH-CTD), family history of PAH (FPAH), modified New York Heart Association (NYHA)/World Health Organization(WHO)functional class I, III, and IV, men aged > 60, renal insufficiency, systolic blood pressure(SBP) < 110 mm Hg, heart rate > 92 beats per min, mean right atrial pressure (mRAP) > 20 mm Hg, 6-minute walking distance(6MWD), brain natriuretic peptide (BNP)> 180 pg/ml, 165 m, brain natriuretic peptide (BNP), 180 pg/mL, pulmonary vascular resistance(PVR)> 32 Wood units, percentage predicted diffusing capacity of lung for carbon monoxide (Dlco) $\leq 32\%$, and presence of pericardial effusion on echocardiogram. Most of the risk factors were associated with increasing mortality rate (indicated by positive sign in β in Table 1), while only four factors were associated with increased one-year survival (indicated by negative sign in β in Table 1).

Risk factors X_i	β	$exp(\beta)$
APAH-CTD	0.7737	1.59
FPAH	1.2801	3.60
APAH-PoPH	0.4624	2.17
Male >60 years age	0.7779	2.18
Renal insufficiency	0.6422	1.90
NYHA Class I	-0.8740	0.42
NYHA Class III	0.3454	1.41
NYHA Class IV	1.1402	3.13
SBP <110 mmHg	0.5128	1.67
Heart Rate >92bmp	0.3322	1.39
6MWD ≥440 m	-0.5455	0.58
6MWD <165 m	0.5210	1.68
BNP <50 pg/ML	-0.6922	0.50
BNP >180 pg/ML	0.6791	1.97
Pericardial effusion	0.3014	1.35
% Dlco ≥80%	-0.5317	0.59
% Dlco ≤32%	0.3756	1.46
mRAP > 20 mmHg	0.5816	1.79
PVR >32 Wood units%	1.4062	4.08

Table 1: A list of 19 binary risk factors, their corresponding coefficients β, and hazard ratio $exp(\beta)$ reported for the PAH REVEAL system (Benza et al., 2010).

To be able to summarize from the model, patients were stratified into five risk groups according to their range of survival probability (Benza et al., 2010) including the low risk group where the predicted 1-year survival probability $> 95\%$, average risk with 90% to 95% survivals, moderately high risk with 85% to 90% survivals, high risk with 70% to 85% survival, and very high risk group with survival probability $< 70\%$.

PAH Calculator

Based on the CPH model, the further application of the CPH model is in the form of a risk calculator. This simplified calculator are useful in everyday clinical practice by helping physicians to decide patient therapies based on level of risk (Benza et al., 2012). The calculator was designed from assigning score to variables according to their hazard ratio. For the risk factors associated with increasing mortality (positive β coefficients), score of two points were assigned for the risk factors which has their hazard ratio ($exp(\beta)$) at least two or more folds, i.e., those with $exp(\beta) \geq 2$, and one point were assigned for other risk factors. Risk factors associated with decreasing mortality (negative β coefficients) were assigned a negative score. Figure 1 shows all risk factors and the interpretation of their hazard ratio rate.

Figure 2 shows the user interface of the PAH Risk Calculator. Each risk factor from the CPH model is listed and mapped with the score. The calculator allows for adding and subtracting the score based on the data entered for an individual patient case. To avoid a negative total score, the base score of 6 is set as a starting score. The total score is interpreted in the same way as the survival probability

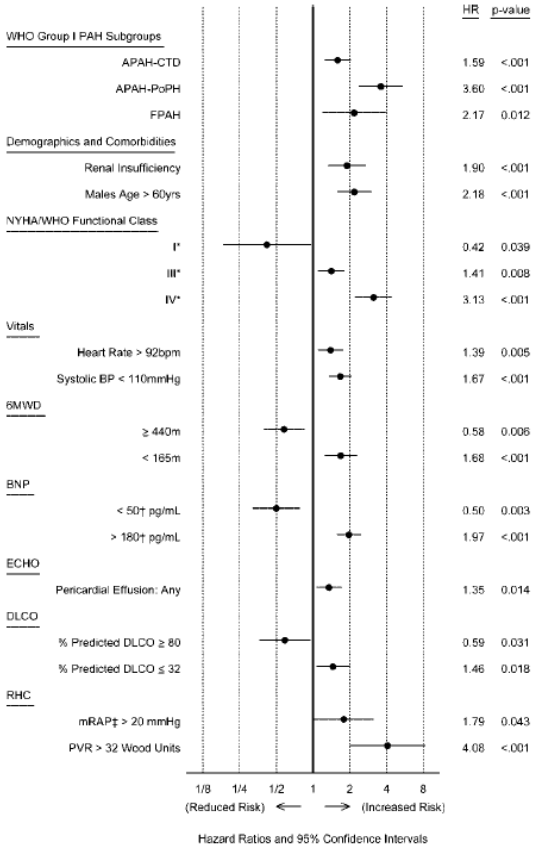


Figure 1: Cox's proportional-hazards of 1-year PAH patients survival variables (Benza et al., 2010) indicating increasing/decreasing mortality rate for each risk factor

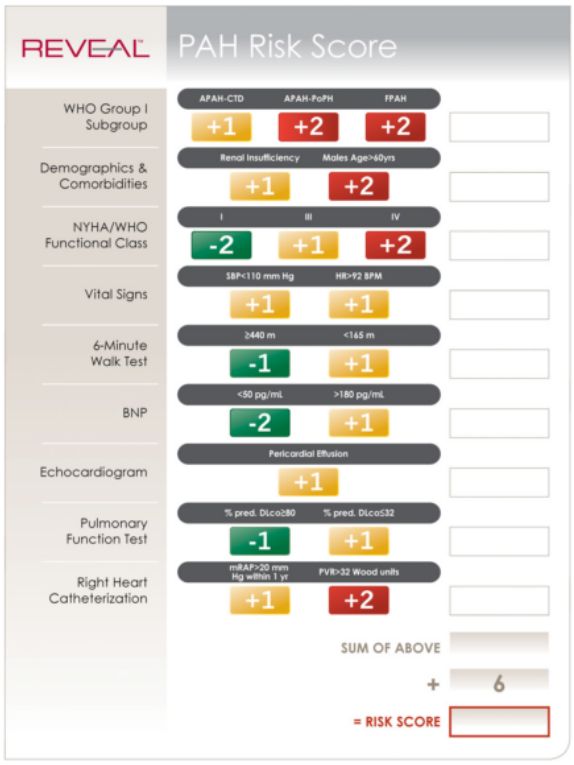


Figure 2: PAH risk score calculator (Benza et al., 2012) (electronic version developed by the United Therapeutics Europe Limited)

given by the CPH model, i.e., it includes the low risk group with the score ≤ 7, average risk with score $= 8$, moderately high risk score $= 9$, high risk with score between 10 and 11, and very high risk group with score ≥ 12. The score, defined as above, makes it simpler for health care providers to use than probabilities.

3 Application of Bayesian Networks to Risk Calculation

An alternative approach to the traditional survival analysis is the use of Bayesian networks (Pearl, 1988) to estimate risks. Compared to the CPH model and several other Artificial Intelligence and Machine Learning techniques, a Bayesian network can model explicitly the structure of the relationships among explanatory variables with their probability (Hanna and Lucas, 2001). A Bayesian network can be built from expert knowledge, available data, or combination of both. If there exists a probabilistic interpretation of existing modeling tool, like in case of the CPH model, a BN model can also be an interpretation of the existing model. The structure of a Bayesian network can depict a complex structure of a problem and provide a way to infer posterior conditional probability distributions, useful for prognosis and diagnosis, including medical decision support systems (Husmeier et al., 2005).

To estimate risks using Bayesian network, the prognosis can be created as a static model, i.e., it can predict the survival at a future point in time. For example, the work of Loghmanpour et al. (2015) focuses on risk assessment models for patients with the left ventricular assist devices (LVADs). Bayesian network have been shown to estimate the risk at various points in time (including 30 days, 90 days, 6 months, 1 year, and 2 years) with accuracy higher than traditional score-based methods (Loghmanpour et al., 2015). An alternative, more complex approach could use dynamic Bayesian networks (DBN), which are an extension of Bayesian networks modeling time explicitly. van Gerven et al. (2007) implemented a DBN for prognosis of patients that suffer from low-grade midgut carcinoid tumor. Instead of treating risk factors independently at each time point, the DBN model considered how the state of patient changed under the influence of choices made by physicians. This model was shown suitable to temporal nature of medical problems throughout the course of care and provide detailed prognostic predictions. However, DBNs requires additional effort during model construction, for example expertise to structure of temporal interaction, large amounts

of (complete) data, which translates to time-consuming efforts (van Gerven et al., 2007).

4 Bayesian Network PAH Risk Calculator

BN Cox model

With no access to the REVEAL Registry data, we created a Bayesian network model that is a formal interpretation of the CPH model, for which the parameters are reported in the literature (Benza et al., 2010). To this effect, we used the method proposed by Kraisangka and Druzdzel (2014). We first created a Bayesian network structure by using all risk factors of the PAH CPH models. We converted all binary risk factors to random variables, which were the parents of the *survival* node. In our case, we have omitted the *time* variable, as the purpose of the PAH Risk Calculator is to capture the risk at one point in time (in this case, it is one year). Figure 3 shows the structure of the BNCox model for the BN-based calculator.

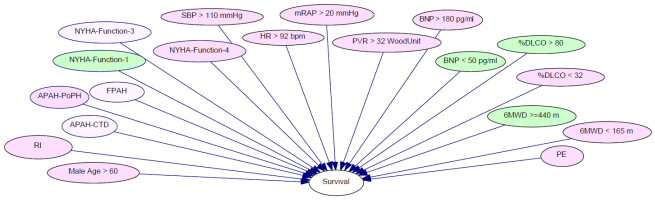


Figure 3: A Bayesian network representing the interaction among variables for the PAH CPH model. All random variables are from the original PAH CPH model and the *Survival* node was added to capture the survival probabilities from the CPH model.

In the next step, we created the conditional probability table for the survival node. The survival probabilities from a CPH model can be encoded into the conditional probabilities as

$$Pr(s \mid X_i, T = t) = S_0\,(t)^{e^{(\beta' x_i)}}\ , \qquad (6)$$

where s means the state of *survived* in the survival node, X_i are all risk factors, T is the time point which is 1 in this case.

We configured all risk factors cases (all binary risk factors generated 2^{19} cases) and obtained all survival probabilities filled in the conditional probability table of a *survival* node. This allowed us to reproduce fully the PAH CPH model by means of a Bayesian network.

BN Interpretation of the PAH Calculator

The original PAH Risk Calculator uses the hazard ratios in the CPH model to derive the risk score for the calculator (Benza et al., 2012). We apply the same approach in our model. Equation 6 captures the survival probabilities s given the states of risk factor. We can extract a hazard ratio of each variable by configuring states of other risk factors to be absent. For example, the hazard ratio of a risk factor x_j can be estimated from

$$\gamma = \frac{\log(Pr(s \mid \bar{x}_1, \ldots, \bar{x}_{j-1}, \mathbf{x_j}, \bar{x}_{j+1}, \ldots, \bar{x}_n))}{\log(Pr(s \mid \bar{x}_1, \ldots, \bar{x}_{j-1}, \bar{\mathbf{x}}_\mathbf{j}, \bar{x}_{j+1}, \ldots, \bar{x}_n))}\ . \qquad (7)$$

The term $\log(Pr(s \mid \bar{x}_1, \ldots, \bar{x}_{j-1}, \bar{\mathbf{x}}_\mathbf{j}, \bar{x}_{j+1}, \ldots, \bar{x}_n))$ is similar to the baseline survival probability in the CPH model ($S_0(1) = 0.9698$). Hence, with this equation, we can track back all hazard ratios.

We use the same criteria as the original PAH Risk Calculator to convert the hazard rate to the score, i.e., score of 2 indicates at least two-fold increase in risk of mortality compared to the baseline risk.

Figure 4 shows a screen shot of our prototype of the Bayesian network risk calculator. The left-hand pane allows for entering risk factors for a given patient case. The right-hand pane shows the calculated score and survival probabilities. Currently, our calculator is a Windows app running on a local server. The numerical risks that produced by the BN calculator are identical to those of the original CPH-based PAH Risk Calculator (Benza et al., 2012).

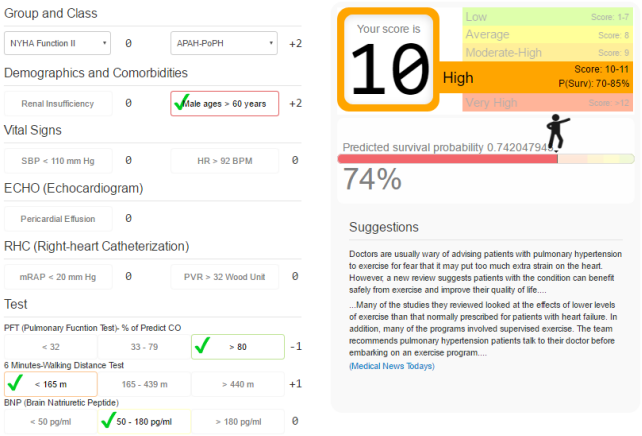


Figure 4: A prototype for Bayesian network risk score calculator for a 1-year PAH prognosis model. The left-hand pane allows for entering risk factors for a given patient case. The right-hand pane shows the calculated score and survival probabilities.

5 Conclusions and Future Work

In this paper, we propose an alternative the the existing Pulmonary Arterial Hypertension (PAH) Risk Calculator that replaces the original Cox Proportional Hazard (CPH) model with a Bayesian network. Because we did not have access to the REVEAL Registry data, we created a Bayesian network model that uses the CPH parameters

learned from the REVEAL Registry data and available in the literature. To this effect, we used a Bayesian network interpretation of the CPH model (Kraisangka and Druzdzel, 2014).

Our calculator reproduces the results of the current PAH Risk Calculator exactly. From this point of view, we have not yet offered a superior calculator. However, we plan to refine the calculator by (1) learning the parameters of the BN model from the data captured in the REVEAL Registry, and (2) enhancing the resulting BN model with medical expert knowledge. The extended model will relax the assumption of the multiplicative character of interactions between the risk factors and the survival variable. It will also relax the assumption that the risk ratio is constant over time. Another direction of our work is allowing risk variables that are not binary. Instead of having 19 binary risk factors, we will be able to group those risk factors that are mutually exclusive, e.g., WHO Group or NYHA/WHO Functional Class. As a result, we can control the number of risk factors and reduce complexities of the model. Yet another direction is allowing dependencies between the risk factors, something that is not straightforward in the CPH model. We should be able to refine the Bayesian network model by using expert knowledge or by training its elements from available data. The current calculator produces a patient-specific score based on hazard ratio. Because the new Bayesian network model will no longer use the multiplicative CPH model, we plan to create new risk score criteria based on the probability of survival rather than the hazard ratio. We have little doubt that with some further modeling effort we should be able to obtain a superior calculator in the sense of producing higher accuracy of the risk estimate than the original CPH-based risk calculator.

Acknowledgements

We acknowledge the support the National Institute of Health under grant number U01HL101066-01 and the Faculty of Information and Communication Technology, Mahidol University, Thailand. Implementation of this work is based on GeNIe and SMILE, a Bayesian inference engine available free of charge for academic teaching and research use at http://www.bayesfusion.com/. While we take full responsibility for any remaining errors and shortcomings of this paper, we would like to thank anonymous reviewers for their valuable suggestions.

References

Allison, P. D. (2010). *Survival Analysis Using SAS: A Practical Guide, Second Edition*. SAS Institute Inc., Cary, NA.

Benza, R. L., Gomberg-Maitland, M., Miller, D. P., Frost, A., Frantz, R. P., Foreman, A. J., Badesch, D. B., and McGoon, M. D. (2012). The REVEAL registry risk score calculator in patients newly diagnosed with pulmonary arterial hypertension. *Chest*, 141(2):354–362.

Benza, R. L., Miller, D. P., Gomberg-Maitland, M., Frantz, R. P., Foreman, A. J., Coffey, C. S., Frost, A., Barst, R. J., Badesch, D. B., Elliott, C. G., Liou, T. G., and McGoon, M. D. (2010). Predicting survival in pulmonary arterial hypertension: Insights from the Registry to Evaluate Early and Long-term Pulmonary Arterial Hypertension disease management (REVEAL). *Circulation*, 122(2):164–172.

Casea, L. D., Kimmickb, G., Pasketta, E. D., Lohmana, K., and Tucker, R. (2002). Interpreting measures of treatment effect in cancer clinical trials. *The Oncologist*, 7(3):181–187.

Cox, D. R. (1972). Regression models and life-tables. *Journal of the Royal Statistical Society. Series B (Methodological)*, 34(2):187–220.

Hanna, A. A. and Lucas, P. J. (2001). Prognostic models in medicine- AI and statistical approaches. *Method Inform Med*, 40:1–5.

Hayes, G. B. (2013). *Pulmonary Hypertension: A Patient's Survival Guide - Fifth Edition, 2013 Revision*. Pulmonary Hypertension Association.

Husmeier, D., Dybowski, R., and Roberts, S. (2005). *Probabilistic modeling in bioinformatics and medical informatics*. Springer.

Irvine, E. J. (2004). Measurement and expression of risk: optimizing decision strategies. *The American Journal of Medicine Supplements*, 117(5):2–7.

Kraisangka, J. and Druzdzel, M. J. (2014). Discrete Bayesian network interpretation of the Coxs proportional hazards model. In van der Gaag, L. C. and Feelders, A. J., editors, *Probabilistic Graphical Models*, volume 8754 of *Lecture Notes in Computer Science*, pages 238–253. Springer International Publishing.

Loghmanpour, N. A., Kanwar, M. K., Druzdzel, M. J., Benza, R. L., Murali, S., and Antaki, J. F. (2015). A new bayesian network-based risk stratification model for prediction of short-term and long-term lvad mortality. *ASAIO Journal*, 61(3):313–323.

Pearl, J. (1988). *Probabilistic reasoning in intelligent systems: networks of plausible inference*. Morgan Kaufmann Publishers Inc., San Francisco, CA, USA.

Subias, P. E., Mir, J. A. B., and Suberviola, V. (2010). Current diagnostic and prognostic assessment of pulmonary hypertension. *Revista Española de Cardiología (English Edition)*, 63(5):583–596.

van Gerven, M. A., Taal, B. G., and Lucas, P. J. (2007). Dynamic Bayesian networks as prognostic models for clinical patient management. *Journal of Biomedical Informatics*, 41(4):515–529.

Measuring the Risk of Public Contracts Using Bayesian Classifiers

Leonardo J. Sales[1,3] **and Rommel N. Carvalho**[1,2]
[1]Department of Research and Strategic Information at the Brazilian Office of the Comptroller General *
[2]Department of Computer Science at the University of Brasília †
[3]Department of Economics at the University of Brasília ‡
{leonardo.sales,rommel.carvalho}@cgu.gov.br

Abstract

Bayesian Classifiers are widely used in machine learning supervised models where there is a reasonable reliability in the dependent variable. This work aims to create a risk measurement model of companies that negotiate with the government using indicators grouped into four risk dimensions: operational capacity, history of penalties and findings, bidding profile, and political ties. It is expected that this model contributes to the selection of contracts to be audited under the central unit of internal control of the Brazilian government, responsible for auditing more than 30,000 public contracts per year.

1 INTRODUCTION

Public contracts can be understood as adjustments made between public administration and private sector for the attainment of public interest objectives (Di Pietro, 1999). The contract terms are set by the governmental unit, this being understood as any body or public authority of federal, municipal, or state level.

Government spending coming from public contracts and direct purchases of goods and services account for approximately 19% of the Brazilian GDP in recent years. Data from the Brazilian Institute of Geography and Statistics (IBGE), published in National Accounts Report in the 2015 fourth quarter, quantifies in R$ 1.07 trillion the amount of government consumption expenditure in that year (IBGE, 2016). The bidding and procurement are the institutional means by which consumption materializes, having important role in the search for efficiency and effectiveness of public spending.

Given the huge number of contracts and purchasing processes to audit, this context raises the challenge of acting effectively in the pursuit of management problems, fraud, and corruption. This is the responsibility of the governmental control units, which specially in Brazil has limited resources.

Take the example of the Office of the Comptroller General (CGU), the central unit of internal control of the Brazilian federal government, which is responsible for auditing any transaction that represents federal spending. The CGU should audit both spending conducted directly (by the central units of the ministries) as the ones conducted indirectly (by almost 20,000 decentralized units), including all payments made by any state or municipality that receives federal funds through voluntary transfers (Brazil, 2003). Nevertheless, the CGU has only 1,200 auditors working directly in the oversight of these expenditures.

In this context, a big issue arises involving the need to rationalize the use of auditing capabilities. There is a clear need to optimize the choice of what will be effectively audited, since the complete census is impossible and uneconomical. Acting in a preventive way to avoid future problems is also important since most of the errors found generate irrecoverable damage, such as paralysis of a engineering project or the need to redo it.

Both the rationalization of choices (in a subsequent operation) and the understanding and treatment of vulnerability (in preventive action) can be analyzed within the more general concept of risk assessment. After all, what is sought in both cases is to identify factors or characteristics of purchases or contracts which increase the chance of future problems such as mismanagement or even fraud.

*SAS, Quadra 01, Bloco A, Edificio Darcy Ribeiro Brasilia, DF, Brazil
†Campus Darcy Ribeiro Brasília, DF, Brazil
‡Campus Darcy Ribeiro Brasília, DF, Brazil

Supervised learning models have been used in similar problems in private sector. Financial institutions assess the risk of potential borrowers, among many suitors with different characteristics and history using such models, in this case called credit scoring (Lessmann et al., 2015). Insurance companies also use such statistical models to assign the value of insurance for a certain good. The techniques learn from the transaction history and quantify the weight of certain characteristics in determining the risk of a client or specific process. Thus, the auto insurance company knows that unmarried young men offer more risk than married women with children.

In practice, these models are applications of statistical and computational techniques of regression and classification using databases that have information of transaction history and labeled cases of "success" and "failure" (Friedman et al., 2001). A good condition in the construction of this type of risk analysis model is the existence of information on transaction history, with variables representing different characteristics of each transaction. Thus, one can distinguish and identify correlations between groups.

This paper proposes to create a predictive model of risk in contracts based on Bayesian classifiers. It will result in the quantification of the propensity that a supplier has problems in government contracts, according to the company's characteristics. Learning models using Bayesian networks are especially useful when you need to organize or discover the knowledge of a particular area through the construction of cause and effect relationships captured from a set of data (Spiegelhalter et al., 1993). Besides this, Bayesian Classifiers have been incorporated into risk measurement studies, especially when it is important to capture and explain the relationships of cause and effect between the different prediction parameters, avoiding the "black box" issue, common in other techniques.

The model will be used to select high-risk contracts to be audited by the CGU and will be based on the estimation of the relations of cause and effect between various indicators that are related to the propensity of contractual risk. The dependent variable is the occurrence of more severe punishment that can be given to a supplier in Brazil: the impediment to bidding. The indicators that will be used as predictors represent characteristics grouped into four risk dimensions: operational capacity, history of penalties and findings, bidding profile, and political ties.

This work is divided into 5 sections. Besides this introduction, Section 2 presents the theoretical framework that supports the central idea of the work and the methodological approach adopted. Section 3 contains the details of the methodology used in the study, including the understanding of data modeling, the creation of the networks, and the validation of the models. Section 4 presents and discusses the results. Finally, Section 5 provides conclusions and considerations on gaps and opportunities for future work.

2 THEORETICAL REFERENCES

In this section we describe the public bidding process in Brazil, the Bayesian classifiers used for learning the predictive models, and some related works.

2.1 PUBLIC BIDDING IN BRAZIL

The whole process of buying products or hiring services in the Brazilian federal government takes place according to the rules of Law 8666/1993 (Brazil, 1993), called Procurement Law. Other regulatory acts complement this law, such as Law 10520/2005 (Brazil, 2002), establishing the types of Auction and Complementary Law 123/2006 (Brazil, 2006) establishing privileges for micro and small businesses in bidding. Law 8666/1993 (Brazil, 1993) details the stages of the bidding process itself, the bidding types allowed, types of contracts, aspects of qualification of companies, and also defines administrative and criminal penalties to be applied to suppliers in case of noncompliance.

The Procurement Law, together with other mentioned legislation, defines the following administrative penalties to suppliers, due to total or partial non-performance of contracts:

- warning;
- pecuniary penalty;
- temporary suspension of bid;
- declaration of non-trustworthiness; and
- impediment to bid and hire.

The whole process of procurement and contracting in the federal government is done using the government's General Services Administration System (SIASG). Each purchase or contract is recorded in this system, since the opening of the process to the issue of commitment.

Existing since 1994, the SIASG started to be used by the government gradually and it already has more than 5 million purchases. All federal administration is required to use this system. Annually it records over 700,000 bids. Some of these bids representing continued provision of services or delivery of goods turns into contracts, generating nearly 30,000 new contracts per year.

2.2 BAYESIAN CLASSIFIERS MODELS

Since Bayesian networks (BNs) have been successfully used in classification problems – *e.g.*, see (Sahami et al., 1998; Friedman et al., 1997; Goldszmidt et al., 2010; Friedman and Goldszmidt, 1996; Cheng and Greiner, 1999; Ceccon et al., 2014; Ye et al., 2014; Shi et al., 2013) –, we decided to experiment with different BN learning algorithms in order to classify the companies that sell service and goods to the government with high likelihood of noncompliance.

Score-based learning is a popular method for inducing BNs. The main idea is to assign a score to a model based on how well it represents the data set used for learning. Thus, the purpose of the algorithm is to maximize the goodness-of-fit score.

In this work we use standard and well-known Bayesian network classifiers, which are aimed at classification. More specifically, we use two algorithms available in the $bnlearn$ R package[1] (Scutari, 2009):

- Naïve Bayes ($naive.bayes$): a simple algorithm that assumes that all explanatory variables are independent of each other. In other words, the target variable is the only parent of all other variables.
- Tree-Augmented Naïve Bayes ($tree.bayes$): an algorithm that relaxes the simple Naïve Bayes assumption of independence, by allowing the explanatory variables to have one other variable as parent besides the target one.

Besides that, we also tried two different score-based learning algorithms, which are also available in the $bnlearn$ R package used in this work (Scutari, 2009):

- Hill-Climbing (hc): a hill climbing greedy search on the space of the directed graphs.
- Tabu Search ($tabu$): a modified hill-climbing able to escape local optima.

The $bnlearn$ package implements random restart with configurable perturbing operations for both algorithms.

A number of different scores were used to fine tune the models learned from the score-based algorithms and to improve their performance, which are also available in the $bnlearn$ package (Scutari, 2009):

- the Akaike Information Criterion score (aic);
- the Bayesian Information Criterion score (bic);
- the logarithm of the Bayesian Dirichlet equivalent score (bde); and
- the logarithm of the modified Bayesian Dirichlet equivalent score ($mbde$).

[1]The package is available at http://www.bnlearn.com/.

2.3 RELATED WORKS

Many studies use supervised learning models in order to predict risk in business transactions. The area where it is more common this type of approach is the bank credit (Lessmann et al., 2015; Hand and Henley, 1997).

These learning models attempt to quantify how the characteristics of potential borrowers influence the probability of default. Classically, the techniques most used for this purpose are Logistic Regression and Discriminant Analysis (Ghodselahi, 2011). Other studies have been testing and comparing some modern techniques (Baesens et al., 2002). In other areas, such as insurance, such models are also widely used.

Bayesian Classifiers have been incorporated into these studies, especially when you want to capture and explain the relationships of cause and effect between the different prediction parameters, avoiding the "black box" issue, common in other techniques (Jiang and Wu, 2009; Zonneveldt et al., 2010; Baesens et al., 2002). Bayesian algorithms provide more clear insights when modeling causal relationships.

A new approach to credit scoring by synthesizing Simple Naïve Bayesian Classifier (SNBC) and the Rough Set Theory is presented by (Jiang and Wu, 2009). A comparison between Naïve Bayes (NB) models, different augmented NB models, and a handcrafted causal network is made by (Zonneveldt et al., 2010).

In the context of public procurement, some initiatives already exist in order to implement similar models in predicting irregularities or contractual problems. For example, Naïve Bayes algorithms are used by (Balaniuk et al., 2012) in an unsupervised approach to quantify the combined risk of private companies and government units in the execution of contracts.

(Sales, 2014) built a model with the same objective of this work (to measure the risk of public contracts) and with similar data. In that case the accuracy using Logistic Regression and Decision Tree were compared, resulting in the best accuracy of 64%.

3 METHODS AND PROCEDURES

The first step in building the Bayesian classification model was the definition of the criteria for characterization of the companies with the highest risk (the "Bad").

In this sense, we chose to characterize the "Bad" group all companies that suffered the following punishments in the years 2015 and 2016: temporary suspension of bid, declaration of non-trustworthiness, and impediment to bid and hire. The group of low-risk companies (hereinafter "Good") are companies with existing contracts in the same period but without such punishment.

The database used contained 1,448 companies, of which 724 were previously classified as "Bad" and other 724 previously classified as "Good"[2].

From this initial setting, the second step was the creation of risk indicators, which cover the past of relations between companies and government, considering the period since 2011, as well as other information that are independent of the period, such as those from the registry of companies. The idea is to answer the following question: What happened in the recent past of the companies that contributed to its contractual default in 2015 and 2016?

These indicators were obtained from the four dimensions of risk: operational capacity, history of penalties and findings, bidding profile, and political ties. The meaning of each of the risk dimensions and some indicators used are described below:

- Operational capacity: irregularities related to the existence or insufficient physical and operational structure of the contracted company.
  - Quantity of indicators: 11.
  - Examples of indicators: number of employees, number of partners, the total amount received from the government, amount received from the government per employee, value received from the government for partner, average salary of employees, average salary of the partners, company size, number of activities carried out by the company, age from the company.
- History of penalties and findings: pre-existence of punishment or audit findings related to the company.
  - Quantity of indicators: 04.
  - Examples of indicators: quantity of received punishments, number of alerts generated in CGU monitoring.
- Bidding profile: company profile when participating in bids, as the average quantity of offers, and the degree of success of business (percentage of wins).
  - Quantity of indicators: 12.
  - Examples of indicators: quantity of purchases, purchase quantity of items, average amount of offers, number of units of the federation, number of wins, percentage of victory, value of contracts, the difference in days between the opening of the company and the first participation in a public procurement.
- Political ties: company relationship with politicians, via donations in campaigns.
  - Quantity of indicators: 01.
  - Examples of indicators: amount donated in political campaigns.

The next step was the transformation of all variables in factors (categories), using a simple process of discretization, where values of each variable were divided into three intervals of equal size. Once complete, the database has been divided in training set (70%) and test (30%). The discretization was carried out due to the limitation of some algorithms used. In future experiments, we will learn models using algorithms that allows continous variables.

At first, we used standard Bayesian classifiers available in the $bnlearn$ R package, Naïve Bayes (NB) and Tree-Augmented Naïve Bayes (TAN).

As the database does not have a very large number of observations, we used a process of estimation with cross-validation in the training subset for both algorithms. The Cross-Validation procedure applied was the random division of training based on 10 sample partitions of equal size, for use in cycles of modeling where 9 partitions are used for training and one for testing. Error measures are then combined to have a single measurement error.

The estimation with cross-validation was performed using a Score-based learning algorithm, which ranks the network structures created with emphasis on model fit. In these algorithms, various parameters can be adjusted in search of the best results forecast.

The loss function used to measure the model results was the misclassification, where the dependent variable value is the result of local distributions (from its parents) and the error function is measured by coincidence or not with the actual values (hit rate).

Since an important aspect of machine learning is the parameter tuning and both NB and TAN in $bnlearn$ do

[2]The 724 companies in the "Bad" group are all companies that meet the criteria described for this class. The 724 companies in "Good" group was obtained by sampling in the set of 41,000 companies that meet the requirements described. Sampling the second group was made in order to solve the dominant class issue, in a process called undersampling (see (Japkowicz et al., 2000) for more details of this process).

not have any parameters to be tuned, we decided to also try another set of algorithms. In $bnlearn$, a set of algorithms that allow many different configurations is the score-based learning algorithms, namely: Hill-Climbing (HC) and Tabu Search (Tabu), both using incremental search. Tabu introduces changes in HC in order to avoid local optima.

In score-based algorithms, it is critical to set the network score calculation method, which measures the quality of the network created using the quantification of posterior probability. Two variables were used in the score parameterization: type of score and penalty parameter. The tested scores types were AIC (Akaike Information Criterion Score), BIC (Bayesian Information Criterion Score), BDE (Bayesian Dirichlet Equivalent Score), and MBDE (Modified Bayesian Dirichlet Equivalent Score), suitable for categorical variables. Besides that, we also tried many different penalty parameters.

The central idea was to try different values of each parameter in order to find the setting that present the best predictive ability. For better understanding, Table 1 shows some of these tested settings and its accuracy measure, aiming to compare the Naïve Bayes (NB) algorithm setting with different configurations[3] of Score-Based algorithms.

Table 1: The table shows that despite our efforts in setting up the Score-Based Algorithms, there was no significant difference than the Naïve Bayes and TAN algorithms. The Accuracy here is the proportion of true results, either true positive or true negative.

Algorithm	Setting	Accuracy - 95% CI
NB	-	(0.70, 0.76)
TAN	-	(0.72, 0.78)
Tabu	AIC, K=0.1	(0.67, 0.73)
Tabu	BDE, ISS=25	(0.65, 0.70)
HC	MBDE, ISS=10	(0.64, 0.70)

4 RESULTS

Since the best models did not present a statistically significant difference in performance and usually the simpler the model the better the generalization, we chose the Naïve Bayes algorithm to run the final model with all the data from the training set in order to check the performance with the test set. The 95% confidence interval of the accuracy was (0.69, 0.77), which shows that the model generalizes well. The sensitivity of the model (prediction ability of "BAD" companies) was 76%. Table 2 shows the results of prediction on the test set.

Table 2: The table shows the model results, that presented a total accuracy of 73%, with higher quality in the identification of "Bad" cases.

	Real values	
Prediction	**Good**	**Bad**
Good	170	47
Bad	69	148
%	71% (specificity)	76% (sensitivity)

We consider this a good result in the context of government contracts, especially when compared with other similar works. Taking as reference the results obtained by (Sales, 2014), you can see a reasonable gain in predictive ability. The sensitivity of the model is particularly important since what really matters is the identification of high-risk cases, even assuming the cost of auditing some low risk contracts, which were misclassified.

5 CONCLUSION AND FUTURE WORK

This work is consistent with a great effort that has been developed by government control institutions to rationalize the use of their human and material resources in order to provide more effective results at lower operating and financial costs.

Considering the current Brazilian context, where a severe economic crisis has been treated through large cuts in public budgets (reducing the sending of resources to control bodies), the efficient use of resources should be a permanent goal.

The attempt to use statistical models based on Bayesian networks is in addition to other initiatives presented in Section 2. The main purpose of these studies is to extract knowledge from various databases that government control institutions have access in order to facilitate the selection of audit objects more likely to present problems.

The classification results are slightly better than other supervised models applied in government databases with the same goal (see (Sales, 2014), described in section 2.3). However, we believe that there is room for improvement in two possible ways: the inclusion of new indicators that capture aspects ignored by this model and the use of optimization algorithms in the parameterization of score-based networks.

[3]The parameters used to set the algorithm were the score-based algorithm, Hill-Climbing (HC) or Tabu Search (Tabu), the score types (AIC, BIC, BDE or MBDE) and the penalty parameter (ISS or K).

Each step in direction of improving these models is a permanent gain for the public auditing activity, and consequently to society.

References

Bart Baesens, Michael Egmont-Petersen, Robert Castelo, and Jan Vanthienen. Learning bayesian network classifiers for credit scoring using markov chain monte carlo search. In *Pattern Recognition, 2002. Proceedings. 16th International Conference on*, volume 3, pages 49–52. IEEE, 2002.

Remis Balaniuk, Pierre Bessiere, Emmanuel Mazer, and Paulo Cobbe. Risk based Government Audit Planning using Nave Bayes Classifiers. In *Advances in Knowledge-Based and Intelligent Information and Engineering Systems*, 2012. URL https://hal.archives-ouvertes.fr/hal-00746198/.

Brazil. Lei n 8666, de 1993, 1993.

Brazil. Lei n 10520, de 2002, 2002.

Brazil. Lei n 10683, de 2003, 2003.

Brazil. Lei Complementar n 123, de 2006, 2006.

S. Ceccon, D.F. Garway-Heath, D.P. Crabb, and A. Tucker. Exploring early glaucoma and the visual field test: Classification and clustering using bayesian networks. *IEEE Journal of Biomedical and Health Informatics*, 18(3):1008–1014, May 2014. ISSN 2168-2194. doi: 10.1109/JBHI.2013.2289367.

Jie Cheng and Russell Greiner. Comparing bayesian network classifiers. In *Proceedings of the Fifteenth Conference on Uncertainty in Artificial Intelligence*, UAI'99, page 101108, San Francisco, CA, USA, 1999. Morgan Kaufmann Publishers Inc. ISBN 1-55860-614-9. URL http://dl.acm.org/citation.cfm?id=2073796.2073808.

Maria Sylvia Zanella Di Pietro. *Direito administrativo*, volume 22. Atlas São Paulo, 1999.

Jerome Friedman, Trevor Hastie, and Robert Tibshirani. *The elements of statistical learning*, volume 1. Springer series in statistics Springer, Berlin, 2001. URL http://statweb.stanford.edu/~tibs/book/preface.ps.

Nir Friedman and Moises Goldszmidt. Building classifiers using bayesian networks. In *Proceedings of the national conference on artificial intelligence*, page 12771284, 1996.

Nir Friedman, Dan Geiger, and Moises Goldszmidt. Bayesian network classifiers. *Machine Learning*, 29 (2-3):131–163, November 1997. ISSN 0885-6125, 1573-0565. doi: 10.1023/A:1007465528199. URL http://link.springer.com/article/10.1023/A%3A1007465528199.

Ahmad Ghodselahi. A hybrid support vector machine ensemble model for credit scoring. *International Journal of Computer Applications*, 17(5):1–5, 2011.

Moises Goldszmidt, James J. Cochran, Louis A. Cox, Pinar Keskinocak, Jeffrey P. Kharoufeh, and J. Cole Smith. Bayesian network classifiers. In *Wiley Encyclopedia of Operations Research and Management Science*. John Wiley & Sons, Inc., 2010. ISBN 9780470400531. URL http://onlinelibrary.wiley.com/doi/10.1002/9780470400531.eorms0099/abstract.

David J Hand and William E Henley. Statistical classification methods in consumer credit scoring: a review. *Journal of the Royal Statistical Society: Series A (Statistics in Society)*, 160(3):523–541, 1997.

IBGE. Indicadores do Instituto Brasileiro de Geografia Estatstica, Contas Nacionais Trimestrais, 2016.

Nathalie Japkowicz et al. Learning from imbalanced data sets: a comparison of various strategies. In *AAAI workshop on learning from imbalanced data sets*, volume 68, pages 10–15. Menlo Park, CA, 2000.

Yi Jiang and Li Hua Wu. Credit scoring model based on simple naive bayesian classifier and a rough set. In *2009 International Conference on Computational Intelligence and Software Engineering*, 2009.

Stefan Lessmann, Bart Baesens, Hsin-Vonn Seow, and Lyn C. Thomas. Benchmarking state-of-the-art classification algorithms for credit scoring: An update of research. *European Journal of Operational Research*, 247(1):124–136, November 2015. ISSN 03772217. doi: 10.1016/j.ejor.2015.05.030. URL http://linkinghub.elsevier.com/retrieve/pii/S0377221715004208.

Mehran Sahami, Susan Dumais, David Heckerman, and Eric Horvitz. A bayesian approach to filtering junk e-mail. In *Learning for Text Categorization: Papers from the 1998 workshop*, volume 62, page 98105, 1998.

Leonardo Jorge Sales. Risk prevention brazilian government contracts using credit scoring. In *Interdisciplinary Insights on Fraud*, chapter 11, pages 264–286. Cambridge Scholars Publishing, 2014.

Marco Scutari. Learning bayesian networks with the bnlearn r package. *arXiv preprint arXiv:0908.3817*, 2009.

Wei Shi, Yao Wu Pei, Liang Sun, Jian Guo Wang, and Shao Qing Ren. The defect identification of LED chips based on bayesian classifier. *Applied Mechanics*

and Materials, 333-335:1564–1568, July 2013. ISSN 1662-7482. doi: 10.4028/www.scientific.net/AMM.333-335.1564. URL http://www.scientific.net/AMM.333-335.1564.

David J. Spiegelhalter, A. Philip Dawid, Steffen L. Lauritzen, and Robert G. Cowell. Bayesian Analysis in Expert Systems. *Statistical Science*, 8(3):219–247, 1993. URL http://www.jstor.org/stable/2245959.

Ye Ye, Fuchiang (Rich) Tsui, Michael Wagner, Jeremy U. Espino, and Qi Li. Influenza detection from emergency department reports using natural language processing and bayesian network classifiers. *Journal of the American Medical Informatics Association*, pages amiajnl–2013–001934, January 2014. ISSN , 1527-974X. doi: 10.1136/amiajnl-2013-001934. URL http://jamia.bmj.com/content/early/2014/01/09/amiajnl-2013-001934.

S Zonneveldt, K Korb, and A Nicholson. Bayesian network classifiers for the german credit data. Technical report, Technical report, 2010/1, Bayesian Intelligence. http://www. Bayesian-intelligence. com/publications. php, 2010.

Bayesian Networks on Income Tax Audit Selection - A Case Study of Brazilian Tax Administration

Leon Sólon da Silva [*]
Secretariat of Federal Revenue of Brazil
Universidade de Brasília
leon.silva@rfb.gov.br

Henrique de C. Rigitano [†]
Secretariat of Federal Revenue of Brazil
henrique.rigitano@rfb.gov.br

Rommel N. Carvalho
Brazil's Office of the Comptroller General [‡]
Universidade de Brasília [§]
rommel.carvalho@cgu.gov.br

João Carlos F. Souza [¶]
Universidade de Brasília
jocafs@unb.br

Abstract

Tax administrations in most countries have more corporate and personal information than any other government office. Data mining techniques can be used in many different problems due to the large amount of tax returns received every year. In the present work we show an essay of the Brazilian Tax Administration on using Bayesian networks to predict taxpayers behavior based on historical analysis of income tax compliance. More specifically, we tried to improve a previous risk based audit selection which detects a large amount of taxpayers as high risk. However, in its current form it identifies much more cases than the tax auditors can handle. Our first results are promising, considerably improving tax audit performance.

1 INTRODUCTION

Tax administrations have more information on people and companies than any other government office. Tax returns, bank transactions, and invoices arrive as hundreds of millions of records every year. The Secretariat of Federal Revenue of Brazil (RFB) is the Brazilian Tax Administration and Brazilian Customs as well. This combination is a major leverage and also a challenge.

Basically, there are two types of taxes: sales taxes and income taxes. Sales taxes includes value-added taxes and they are based on the value of the product being sold. Income tax is based on how much a person or a company earns. In most countries, sales taxes amount are considerably larger than income taxes (OECD, 2013). In Brazil, corporate and personal income taxes are about 50% of the country's revenue (RFB, 2016). Although corporate tax has much greater impact on final numbers, personal income tax audits affects a considerably large share of the Brazilian citizens. There are 27 million individual taxpayers in Brazil, about 13% of the population (RFB, 2016).

In order to facilitate and prioritize tax audits on personal income tax, RFB created the concept of a "fiscal lattice". One can understand the fiscal lattice as a first audit selection based on historical risk analysis of tax compliance by taxpayers. This lattice is a complex process in which many tax auditors specialized in personal income tax frauds create risk based rules for audit selection. The main difference between a regular audit and fiscal lattice audit is that the former has a much simpler process of analysis in order to determine whether to punish a taxpayer or not.

Since the number of taxpayers has increased, and the ratio between tax auditors and citizens has been reducing (RFB, 2016), the number of income taxpayers caught on fiscal lattice has increased as well. From 2010 to 2014, the taxpayers selected for this kind of audit highly increased (RFB, 2016). This changing scenario is pushing the tax administration to a limit of the tax auditors capacity of analysis. RFB's major office, has about 10,000 tax auditors and a huge backlog of fiscal lattice audits to analyze.

Data mining techniques can help better selecting taxpayers for audit and the present work offers one solution to improve the selection of this kind of audits. In Section 2.1 we discuss how Bayesian networks can be used as a classification algorithm in order to create predictive models.

The document is organized as follows: Section 2 describes some background information about Bayesian

[*] Anexo Ministério da Defesa, 5o andar Brasília, DF, Brazil
[†] Av. Rogerio Weber, 1752 - Centro, Porto Velho, RO, Brazil
[‡] SAS, Quadra 01, Bloco A, Edificio Darcy Ribeiro Brasilia, DF, Brazil
[§] Campus Darcy Ribeiro Brasilia, DF, Brazil
[¶] Campus Darcy Ribeiro Brasilia, DF, Brazil

networks; Section 3 details the solution for the tax audit selection problem, from its methodology to our first results; Section 4 presents the conclusion and future work.

2 BACKGROUND

In this section we bring some tax administration concepts, formulate the problem assessed by the present work, and discuss Bayesian networks for prediction.

2.1 BAYESIAN NETWORKS FOR PREDICTIVE MODELS

As stated by (Korb and Nicholson, 2010) Bayesian networks (BNs) are graphical models for reasoning under uncertainty, where the nodes represent variables (discrete or continuous) and arcs represent direct connections between them. These direct connections are often causal connections. In addition, BNs model the quantitative strength of the connections between variables, allowing probabilistic beliefs about them to be updated automatically as new information becomes available.

Bayesian networks are useful to learn from data and discover causalities between variables and it can be used as a classifier algorithm. It is being used for prediction in many different problems, from genetics (Jansen et al., 2003) and prognostics of breast cancer (Gevaert et al., 2006), to identification of split purchases (Carvalho et al., 2014). In the present work, we use Bayesian networks as a solution for predicting a taxpayer to be compliant or non-compliant in terms of tax obligations. In more detail, our approach presents an improvement of tax audit selection using Bayesian networks to build predictive models. In the next section we present the details for the solution to our problem, as well as the first results.

The next subsections describe two different types of Bayesian networks, Naïve Bayes and Tree-Augmented Naïve Bayes.

2.1.1 Naïve Bayes

Naïve Bayes is the most simple version of Bayesian network. It uses strong connections between the nodes and it considers all explanatory variables (nodes) as independent. Despite its simpleness it has many applications with good results and great run performance as stated in (Zhang, 2004).

2.1.2 Tree-Augmented Naïve Bayes

Tree-Augmented Naïve Bayes (TAN), as explained in (Zheng and Webb, 2011), relaxes the assumption of complete independence of the explanatory variables by enforcing a tree structure. In this case, each explanatory variable only depends on the class and one other variable. This relaxation allows the representation of more complex models, leading to possible performance improvements, as shown in (Carvalho et al., 2014).

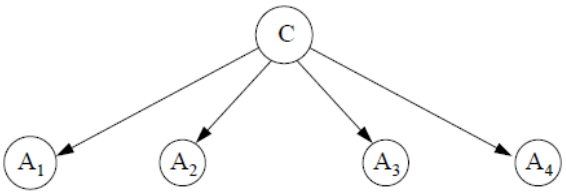


Figure 1: Example of Naïve Bayes Network (Zhang, 2004)

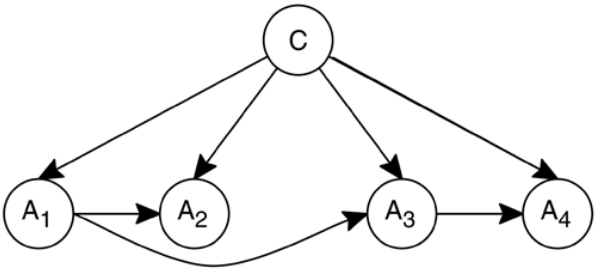


Figure 2: Example of Tree-Augmented Naïve Bayes Network (Jiang et al., 2009)

2.2 RELATED WORK

As stated in (Silva et al., 2015) many tax administrations have been using data mining techniques to create predictive models for tax compliance risk. Despite being a topic of great interest, tax administrations have many concerns in publishing internal projects. Since taxpayer information is classified and should be protected by tax officers, many of them do not share the details of tax compliance risk projects.

A source of such information, case studies, methodolo-

gies, and best practices are intergovernmental organizations. For tax administrations and customs the World Customs Organization (WCO) and the Organization for Economic Cooperation and Development (OECD) are important sources. In a recent survey that gathered many countries, OECD presented a comparative chart that shows the use of data mining to detect tax fraud (OECD, 2013).

Tax Administrations internal publications also present many studies that can be applied by other countries and many of them have developed methodologies based on statistical analysis and data mining to create tax compliance risk systems. Most countries use data mining for taxpayers classification considering its risks of non-compliance.

Some studies, however, reveal different data analysis approach being held in tax administration. The US Internal Revenue Service (IRS) uses data mining for different purposes, according to (Castellón González and Velásquez, 2013), among which are tax compliance risk based taxpayer classification, tax fraud detection, tax refund fraud, criminal activities, and money laundering (Watkins et al., 2003).

Another related reference is Jani Martikainens master thesis (Martikainen et al., 2012). He presents results of studies conducted by the Australian Taxation Office (ATO) concerning the usage of models to detect high-risk tax refund claims. Also according to the author, the ATO avoided the payment of refunds of about US$ 665,000,000.00 between 2010 and 2011 based on data mining tools. ATO uses refund models based on social networking discovery algorithms that detect connections between individuals, companies, partnerships, or tax returns. The models are updated and refined to enhance detection and increase the recognition of new fraud (Martikainen et al., 2012).

More related to the present work Gupta *et al.* in (Gupta and Nagadevara, 2007) describes in details different approaches on using data mining techniques to improve tax audit selection. The main difference is that in (Gupta and Nagadevara, 2007) the main taxes are value-added taxes in contrast with income taxes, object of the present research. Also in (Kirkos et al., 2007) data mining is used to detect frauds on financial statements, which can be easily customized to tax returns and tax evasion/fraud.

3 SOLUTION AND FIRST RESULTS

In this section we describe the methodology used in the present work and detail each step of the data analysis from the information and data gathering to the construction of predictive models for improvement of tax audit selection.

3.1 METHODOLOGY

The methodology of the present work follows the well-known CRISP-DM (CRISP-DM). The Cross Industry Standard Process for Data Mining is a technology-independent methodology and reference model to implement data mining process in every business. It describes each phase every data mining work should pass. Each phase is equally relevant for the success of the data analysis process and should not be underestimated. The process has six phases and it is possible to perform the same step more than once. The phases of CRISP-DM are (Wirth and Hipp, 2000):

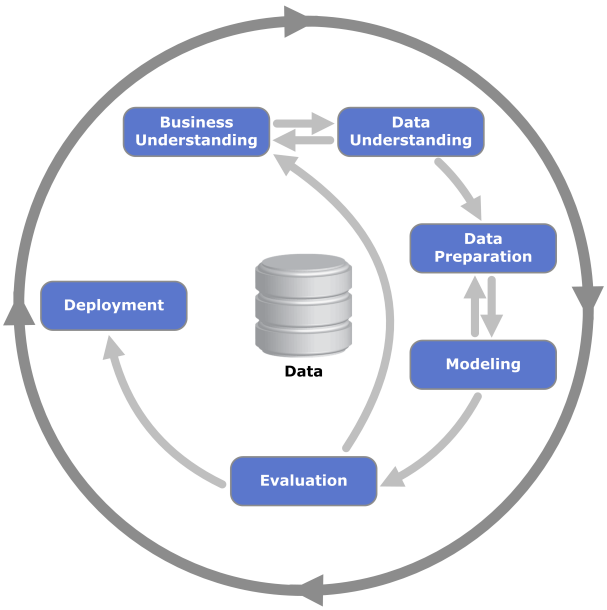


Figure 3: CRISP-DM Reference Model (Wirth and Hipp, 2000)

Business Understanding

Every data analysis process is designed to answer business questions to achieve business goals. In the business understanding phase of CRISP-DM these questions are asked and possible solutions are also proposed. Possible quantitative and qualitative business process' improvements are also detailed, in order to justify the use of data mining techniques to solve business problems.

According to (Chapman et al., 2000), this initial phase focuses on understanding the project objectives and requirements from a business perspective, and then converting this knowledge into a data mining problem definition, and a preliminary project plan designed to achieve the objectives.

Data Understanding

Once the business questions are clear, it is time to understand the required information to perform the changes needed in the business process and achieve the goals identified in the previous phase. In data understanding, all sources of information needed to perform the analysis are determined. The first insights and main patterns are also identified in the first contact with the data available from the possible sources. Each business question needs to be mapped to every data source (systems, databases, webpages, etc.) in order to address every goal and identify possible gaps and lack of information.

In (Wirth and Hipp, 2000) it is stated that there is a close link between business understanding and data understanding. The formulation of the data mining problem and the project plan require at least some understanding of the available data.

Data Preparation

The data preparation phase covers all activities to construct the final dataset (data that will be fed into the modeling tool(s)) from the initial raw data. Data preparation tasks are likely to be performed multiple times, and not in any prescribed order. Tasks include table, record, and attribute selection, data cleaning, construction of new attributes, and transformation of data for modeling tools.

Modeling

In this phase, various modeling techniques are selected and applied, and their parameters are calibrated to optimal values. Typically, there are several techniques for the same data mining problem type. Some techniques require specific data formats. There is a close link between data preparation and modeling. Often, one realizes data problems while modeling or one gets ideas for constructing new data.

Evaluation

At this stage in the project you have built one or more models that appear to have high quality, from a data analysis perspective. Before proceeding to final deployment of the model, it is important to more thoroughly evaluate the model, and review the steps executed to construct the model, to be certain it properly achieves the business objectives. A key objective is to determine if there is some important business issue that has not been sufficiently considered. At the end of this phase, a decision on the use of the data mining results should be reached.

Deployment

Creation of the model is generally not the end of the project. Usually, the knowledge gained will need to be organized and presented in a way that the customer can use it. Depending on the requirements, the deployment phase can be as simple as generating a report or as complex as implementing a repeatable data mining process. In many cases it will be the user, not the data analyst, who will carry out the deployment steps. In any case, it is important to understand up front what actions will need to be carried out in order to actually make use of the created models.

3.2 BUSINESS UNDERSTANDING

Our main goal is to improve individuals tax audit selection. We try to achieve a better audit process performance by better using the tax auditors knowledge and time available to perform these audits. As in any tax administration, there are far more taxpayers returns and information to analyze than tax officers, and to achieve the revenue goals and tax fairness it is major that the selection of audit is as risk based as possible.

In Brazil, personal taxpayers pay their income taxes every month. Since the tax is calculated on a year based, by April of the next year, taxpayers are obliged to send their income tax return in order to adjust their debt (or credit). Every year, tens of million of returns are sent to RFB, much more than it could handle if there were no risk based selection.

RFB created the concept of “fiscal lattice” to select personal income tax returns based on tax compliance risk. In this technique personal income tax fraud experts analyze the historical of all taxpayers and their previous knowledge in order to come up with parameters to select the tax returns for audit. Once caught on “fiscal lattice”, only a tax officer could release the tax return, preventing fraudsters from receiving a possible credit. There are three main purposes in using this technique:

- to better select taxpayers based on tax compliance risk;
- to facilitate the verification of tax auditors, since each parameter has a well defined analysis and treatment activities;
- to ease the auto-correction of tax returns by taxpayers, since many of them were caught due to filling

errors.

Besides all Brazilian tax administration efforts to select the individuals tax audits, the number of audits selected by fiscal lattice has increased from 569,000 in 2011 to 937,000 in 2014[1] in contrast with the number of tax officers, that decreased from 12,273 in 2010 to 10,419 in 2015 (RFB, 2016).

More specifically, we intend to use data mining techniques to discharge as many taxpayers as possible of fiscal lattice, with the minimum compliance risk to tax administration. With thousands of audits already finalized by experienced tax auditors, it is possible to assess this problem with machine learning tools and achieve best results in letting go those taxpayers that offer less risk of tax compliance.

In our first approach on trying data mining techniques to address the problem, we selected a certain RFB's unit that has been suffering from the large number of fiscal lattice audits. The "Delegacia Especial de Pessoa Fsica" (DERPF) or "Individual Taxpayers Special Office" is an individual taxpayer specialized unit located at Sao Paulo City, the Brazilian biggest city, in the most economically active federation unit (State of Sao Paulo). This unit has come to its limit of fiscal audits since its creation in 2014, and has the largest number of this kind of audits in the whole country. It was a natural choice for our first experiments.

3.3 DATA UNDERSTANDING AND PREPARATION

To answer the business question on how to improve the selection of individual taxpayers caught in fiscal lattice, we evaluated the sources of the information needed to perform the data mining analysis. Our sample was taken from audits performed by DERPF from years 2014 to 2016.

Basically, all individuals taxpayer information was taken from internal systems, from online systems to datamarts and datawarehouses. Most of taxpayer information caught in fiscal lattice is available from tax returns, but some information is taken from invoices and financial operations. The exact properties retrieved by the data extraction as well as the fraud/non-compliance rate are classified information.

The final taxpayer table has 25,322 taxpayer's returns analyzed by tax auditors and classified as compliant or non-compliant. Each line has, besides the dependent variable (compliant) other 20 characteristics of taxpayers and information retrieved from returns and other systems. From these, 13,547 are women and 10,730 are men. Other explanatory variables are information of tax return and unfortunately cannot be specified because it could present classified information, since the result of the analysis could lead taxpayers to learn fraud patterns and use that information to avoid being caught.

For preparation, all independent variables were analyzed in order to remove the incomplete rows and to discretize continuous ones to comply with the Bayesian network algorithms constraints. The numeric variables where classified within bands in terms of average multipliers (one average, half average, three times average, etc.). After data preparation the final number of individual taxpayers returns was 24,277.

All data preparation took place using R language[2] and its packages.

3.4 MODELING AND EVALUATION

We used $bnlearn$ R package[3] in order to run the Bayesian network algorithms. Specifically the functions $naive.bayes$ and $tree.bayes$ where chosen to create the predictive models. The first is the well-known Naïve Bayes algorithm, which does not take parameters for customizing the models and the former is an implementation of the Tree-Augmented Naïve Bayes (TAN) algorithm. The TAN algorithm takes white list (force the inclusion of arcs in Bayesian network), black list (force the exclusion of arcs in Bayesian network), and mi[4] parameters.

To create the predictive models we took the compliant variable as dependent and the other 35 (thirty five) information as independent variables. The sample of 24,277 where divided into training (80%) and test (20%). No validation sample was needed since we used 10-fold cross-validation technique with $bnlearn$'s function $bn.cv()$.

As stated in $bn.cv()$ documentation (CRAN, 2016) k-fold is a technique where the data is split in k subsets of equal size. For each subset in turn, bn is fitted (and possibly learned as well) on the other k - 1 subsets and the loss function is then computed using that subset. Loss estimates for each of the k subsets are then combined to give an overall loss for data.

[1] In 2015 this number decreased to 670,000 due to efforts in better selecting individuals tax returns for audits

[2] `https://www.r-project.org/`.

[3] `http://www.bnlearn.com/`.

[4] The estimator used for the mutual information coefficients for the Chow-Liu algorithm in TAN. Possible values are mi (discrete mutual information) and mi-g (Gaussian mutual information). We use discrete since all explanatory variables have been discretized

Since the proportion of compliant/non-compliant taxpayers is classified information, we present the results of the predictive models in terms of improvements from the actual process of discharging taxpayers from fiscal lattice. Since our dependent variable is compliant/non-compliant, we are interested in evaluating the models by specificity more than sensitivity, since it is more dangerous to let a non-compliant taxpayer go away without being audited than to select one that is compliant to be audited.

Each Brazilian tax administration local unit is autonomous and may choose whatever criteria it finds best to dismiss taxpayers from fiscal lattice. So, to a matter of possible comparison with our proposal, we consider a linear cut (random selection) of taxpayers until it reaches a units capacity. If, for example, an office has the capacity to audit 2,000 taxpayers per month, and there are 3,000, we consider the actual process to randomly choose the 1,000 to be dismissed. The overall taxpayers wrongly dismissed, is the same as the proportion between non-compliant taxpayers from overall caught on fiscal lattice. Our goal is to better predict if a taxpayer caught on fiscal lattice is compliant or not. If we come to a specificity considerably better than random selection, we achieve our goal to let go as few non-compliant taxpayers as possible.

As we learn from Table 1, using Naïve Bayes is already a good tool to select those taxpayers which can and cannot be dismissed from being audited. Tree-Augmented Naïve Bayes had no major advantages, despite the customization of parameters (root chose automatically or user defined).

Table 1: Predictive Models by Algorithm/Parameters

Algorithm	Performance Rate
Naive Bayes	41 %
TAN (auto root)	34 %
TAN (selected root)	35 %

Therefore, the predictive models in this first results showed optimistic results, resulting in a increase of more then 30% in tax audit selection in comparison to randomly discharging taxpayers. It is major to recollect that the taxpayers caught in fiscal lattice have already been through a risk based process of selection and any increase in this criteria is a leverage in using Bayesian networks to build models of tax compliance.

4 CONCLUSION AND FUTURE WORK

Brazil has been through a major crisis and the responsibility of the RFB as a tax administration has also increased in order to guarantee the revenue for public policies. A better selection of tax audits save resources and increase the performance of the collecting tax process. Our approach on creating predictive models to improve the risk based selection of the so called "fiscal lattice" proved to be a promising one based on the first results.

We intend to use different approaches and Bayesian networks algorithms in order to create compliance risk scores and leave the decision of taxpayers being compliant or not to the tax officers and possibly increase the specificity. The approach in the present work delegates this decision to the prediction algorithm.

Furthermore we will try and build Bayesian networks with larger samples and more tax units and include more information about the taxpayer, since in this work we basically used income tax returns and registry information. Financial transactions and invoice data could be interesting explanatory variables and will be used in future applications.

Acknowledgements

The authors would like to thank RFB, specially DERPF, for providing the resources necessary to work in this research, as well as for allowing its publication.

References

Rommel N Carvalho, Leonardo Sales, Henrique A Da Rocha, and Gilson Libório Mendes. Using bayesian networks to identify and prevent split purchases in brazil. In *BMA@ UAI*, pages 70–78, 2014.

Pamela Castellón González and Juan D Velásquez. Characterization and detection of taxpayers with false invoices using data mining techniques. *Expert Systems with Applications*, 40(5):1427–1436, 2013.

Pete Chapman, Julian Clinton, Randy Kerber, Thomas Khabaza, Thomas Reinartz, Colin Shearer, and Rudiger Wirth. Crisp-dm 1.0 step-by-step data mining guide. 2000.

CRAN. Cran project. package bnlearn. https://cran.r-project.org/web/packages/bnlearn/index.html, 2016. Accessed: 2016-05-08.

Olivier Gevaert, Frank De Smet, Dirk Timmerman, Yves Moreau, and Bart De Moor. Predicting the prognosis of breast cancer by integrating clinical and microarray

data with bayesian networks. *Bioinformatics*, 22(14): e184–e190, 2006.

Manish Gupta and Vishnuprasad Nagadevara. Audit selection strategy for improving tax compliance: Application of data mining techniques. In *Foundations of Risk-Based Audits. Proceedings of the eleventh International Conference on e-Governance, Hyderabad, India, December*, pages 28–30, 2007.

Ronald Jansen, Haiyuan Yu, Dov Greenbaum, Yuval Kluger, Nevan J Krogan, Sambath Chung, Andrew Emili, Michael Snyder, Jack F Greenblatt, and Mark Gerstein. A bayesian networks approach for predicting protein-protein interactions from genomic data. *Science*, 302(5644):449–453, 2003.

Liangxiao Jiang, Harry Zhang, and Zhihua Cai. A novel bayes model: Hidden naive bayes. *Knowledge and Data Engineering, IEEE Transactions on*, 21(10): 1361–1371, 2009.

Efstathios Kirkos, Charalambos Spathis, and Yannis Manolopoulos. Data mining techniques for the detection of fraudulent financial statements. *Expert Systems with Applications*, 32(4):995–1003, 2007.

Kevin B Korb and Ann E Nicholson. *Bayesian artificial intelligence*. CRC press, 2010.

Jani Martikainen et al. Data mining in tax administration-using analytics to enhance tax compliance. *Department of Information and Service Economy. Aalto University*, 2012.

OECD. Tax administration 2013 - comparative information on oecd and other advanced and emerging economies. Technical Report 2308-7331, Organisation for Economic Co-operation and Development, Paris, 2013. URL `http://www.oecd-ilibrary.org/content/serial/23077727`.

RFB. Secretariat of federal revenue of brazil (rfb) website. http://www.receita.fazenda.gov.br, 2016. Accessed: 2016-05-08.

Leon Sólon da Silva, Rommel Novaes Carvalho, and João Carlos Felix Souza. Predictive models on tax refund claims-essays of data mining in brazilian tax administration. In *Electronic Government and the Information Systems Perspective*, pages 220–228. Springer, 2015.

R CORY Watkins, K Michael Reynolds, Ron Demara, Michael Georgiopoulos, Avelino Gonzalez, and Ron Eaglin. Tracking dirty proceeds: exploring data mining technologies as tools to investigate money laundering. *Police Practice and Research*, 4(2):163–178, 2003.

Rüdiger Wirth and Jochen Hipp. Crisp-dm: Towards a standard process model for data mining. In *Proceedings of the 4th international conference on the practical applications of knowledge discovery and data mining*, pages 29–39. Citeseer, 2000.

Harry Zhang. The optimality of naive bayes. *AA*, 1(2):3, 2004.

Fei Zheng and Geoffrey I Webb. Tree augmented naive bayes. In *Encyclopedia of Machine Learning*, pages 990–991. Springer, 2011.

Target Beliefs for SME-oriented, Bayesian Network-based Modeling

Robert Schrag
Haystax Technology
11210 Corsica Mist Ave
Las Vegas, NV 89135

Edward Wright, Robert Kerr, Robert Johnson
Haystax Technology
8251 Greensboro Dr, Suite 1000
McLean, VA 22102

Abstract

Our framework supporting non-technical subject matter experts' authoring of useful Bayesian networks has presented requirements for fixed probability soft or virtual evidence findings that we refer to as target beliefs. We describe exogenously motivated target belief requirements for model nodes lacking explicit priors and mechanistically motivated requirements induced by logical constraints over nodes that in the framework are strictly binary. Compared to the best published results, our target belief satisfaction methods are competitive in result quality and processing time on much larger problems.

1. INTRODUCTION

The variety of soft or virtual evidence finding on a Bayesian network (BN) node in which a specified probability distribution must be maintained during BN inference—called a fixed probability finding by (Ben Mrad, 2015) and called a target belief here—has received limited attention. Published results for inference algorithms respecting such findings have addressed small, artificial problems including at most 15 nodes (Peng et al., 2010; Zhang et al., 2008).

Our work on one real application has required addressing dozens of such findings in a BN comprising hundreds of nodes. In this context, target beliefs are motivated by modelers' need to address authoritative sources exogenous to the model itself, where beliefs should hold for selected non-BN root model nodes—i.e., nodes lacking explicit prior probability distributions (that otherwise might be used to achieve target beliefs directly).

For example, if a binary node Divorces appears deep in a person risk assessment network as an indicator of a top-level binary node Trustworthy, usually (without target beliefs or other node findings) the network's computed belief in Divorces will depend on the network's conditional probability tables (CPTs)[1]—not on a published statistic about the divorce rate in an intended subject population. To make our model's belief in Divorces agree with the exogenous statistic, a modeler can:

1. Adjust CPTs throughout the model to agree with the exogenous specification.
2. Invoke Jeffrey's rule (Jeffrey, 1983) to compute a likelihood finding on Divorces that achieves the specified belief.
3. Specify a target belief for Divorces and rely on target belief satisfaction machinery to achieve the target.

The first option is not entirely compatible with our modeling framework.[2] The modeler's manual effort under either of the first two options may be undermined as soon as s/he modifies the model again.[3] The last option offloads the work of target belief satisfaction to an automated process—at the expense of executing that process, as often as necessary. Execution time may be acceptable for a given use case if the model is small, if it is not modified often, or if model development is sufficiently simplified under this approach to enhance overall productivity. As we intend our framework to be subject matter expert- (SME-)friendly, this option is attractive. The more we can free a modeler to concentrate on higher-level decisions with greater domain impact, the more and better models s/he should be able to deliver.

[1] Including top-level node priors as a degenerate case.

[2] Our framework automatically computes CPTs (see section 2) to reflect a modeler's specified strength with which a child node (counter-)indicates its parent node. So, modifying CPTs is appropriate only when modifying these strengths is. Likewise, the representation would not naturally accommodate a conventional approach to machine learning of CPT entries.

[3] In principle, any of a large variety of modifications—including more invocations of this option to address additional exogenous probabilities—could affect computed belief in Divorces.

Our work adapting the framework to realize probabilistic argument maps for intelligence analysis (Schrag et al., 2016a; 2016b) has surfaced powerful representations (Logic constraints—see section 4) that can improve model clarity and correctness and that often require target beliefs.

In the following sections, we outline the framework, our large person risk assessment model, and the view of framework models as probabilistic argument maps. We explain how Logic constraints can improve arguments (models) and how target beliefs can support such constraints. We briefly review existing competitive target belief processing methods, then describe our own method and results.

2. SME-ORIENTED MODELING FRAMEWORK

We developed the framework to facilitate creation of useful BNs by non-technical SMEs. Faced with the challenge of operationalizing SMEs' policy-guided reasoning about person trustworthiness in a comprehensive risk model (Schrag et al., 2014), we first developed a model encoding hundreds of policy statements. The need for SMEs both to understand the model and to author its elements inspired us to develop and apply a technical approach using exclusively binary random variables (BN nodes) over the domain {true, false}. This led us to an overall representation that happens to extend standard argument maps (CIA, 2006) with Bayesian probabilistic reasoning (Schrag et al., 2016a; 2016b).

In the framework, every node (or argument map statement[4]) is a Hypothesis. Some Hypotheses are Logic nodes whose CPTs are deterministic. Connecting the nodes are links whose types are listed in Table 1. Argument maps' SupportedBy and RefutedBy links correspond to our IndicatedBy and CounterIndicatedBy links.

Table 1: Framework link types (center column). For the last two link types, the argument map-downstream statement (BN-downstream node) is a Logic node.

Argument map-downstream[5] statement	IndicatedBy	Argument map-upstream statement(s)
	CounterIndicatedBy	
	MitigatedBy	
	RelevantIf	
	OppositeOf	
	ImpliedByConjunction	
	ImpliedByDisjunction	

We encode strengths for non-Logic node-input links (first four rows of Table 1) using fixed odds ratios per Figure 1.

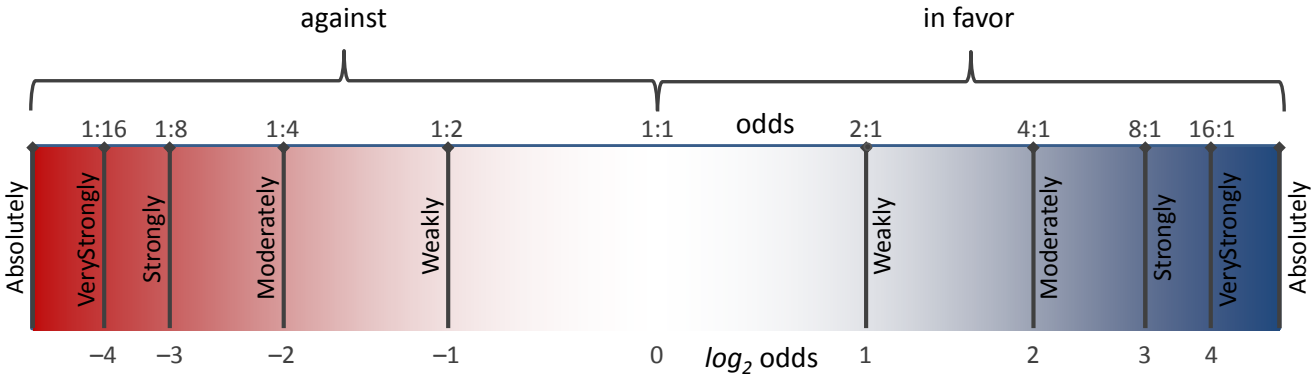


Figure 1: Odds ratios for discrete link strengths. Absolutely is intended as logical implication. We do not otherwise commit SMEs to absolute certainty.

[4] Our binary BN nodes correspond to propositions bearing truth values. In the argument map point of view, these propositions may be understood to be statements.

[5] Per argument map convention, "downstream" is left, "upstream" right in the left-flowing argument map of Figure 3. Except for Logic nodes, this is opposite of links' causal direction in BNs.

A framework process (Wright et al., 2015) converts specifications into corresponding BNs. The conversion process recognizes a pattern of link types incident on a given node and constructs an appropriate CPT reflecting specified polarities and strengths. The SME thus works in a graphical user interface (GUI) with an argument map representation (as if at a "dashboard"), and BN mechanics and minutiae all remain conveniently "under the hood."

The framework includes stock noisyOr and noisyAnd distributions (bearing a standard Leak parameter) for BN nodes with more than one parent. While these have so far been sufficient in our modeling efforts, we also could fall back to fine distribution specification. We have deliberately designed the framework to skirt standard CPT elicitation, which can tend to fatigue SMEs. Consider an indicator of h different Hypotheses, so with h BN parents and 2^h CPT rows. Suppose belief is discretized on a 7-point scale.[6] Then standard, row-by-row elicitation requires 2^h entries. With noisyOr or noisyAnd, we need only h entries bearing a polarity and strength for each parent, plus a Leak value for the distribution.

We are working to make modeling in the framework more accessible to SMEs, particularly via model editing capabilities in the GUI exhibited in Figure 3. (Schrag et al., 2016a) describes our framework encoding of an analyst's argument, favorable comparison of resulting modeled probabilities to analyst-computed ones, and favorable comparison of CPTs generated by the framework vs. elicited directly from analysts.

3. PERSON RISK MODEL WITH EXOGENOUS BELIEF REQUIREMENTS

Our person risk assessment application includes a core generic person BN accounting for interactions among beliefs about random variables representing different person attribute concepts like those in Figure 2.

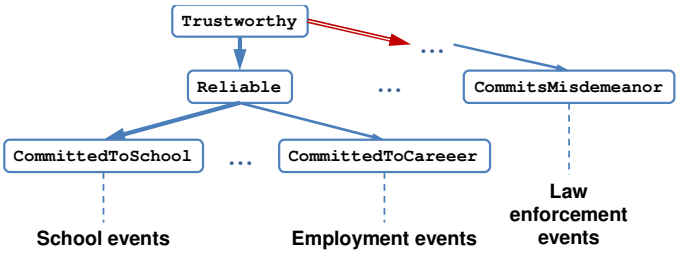


Figure 2: Partial generic person attribute concept BN (top), with related event categories (bottom). BN influences point (causally) from indicated concept hypothesis to indicating concept. Stronger indications have thicker arrows. A single negative indication has a red, double-lined arrow.

The framework processes a given person's event evidence to specialize this generic BN into a person-specific BN (Schrag et al., 2014).

We have specified target beliefs for some two dozen nodes in the generic person network. By processing the target beliefs in an event evidence-free context, we ensure that events have the effects intended, respecting both indication strengths and exogenous statistics.[7]

4. INTELLIGENCE ANALYSIS MODEL MOTIVATING REQUIRMENTS FROM LOGIC CONSTRAINTS

Figure 3 is a screenshot of a model addressing the CIA's Iraq retaliation scenario (Heuer, 2013)[8], where Iraq might respond to US forces' bombing of its intelligence headquarters by conducting major, minor, or no terror attacks, given limited evidence about Saddam Hussein's disposition and public statements, Iraq's historical responses, and the status of Iraq's national security apparatus. This model emphasizes Saddam's incentives to act. By setting a hard finding of false on the incentive-collecting node SaddamWins, we can examine computed beliefs under Saddam's worst-case scenario (and, by comparing this to his best-case scenario, determine that conducting major terror attacks is not his best move). See (Schrag et al., 2016a) for details.

[6] As (Karvetski et al., 2013) note, the inference quality of models developed this way usually rivals that of models developed with arbitrary-precision CPTs.

[7] Such a dividing line between generic model and evidence may not be so bright in a probabilistic argument map, where an intelligence analyst may enter both hypothesis and evidence nodes incrementally.

[8] See chapter 8, "Analysis of Competing Hypotheses."

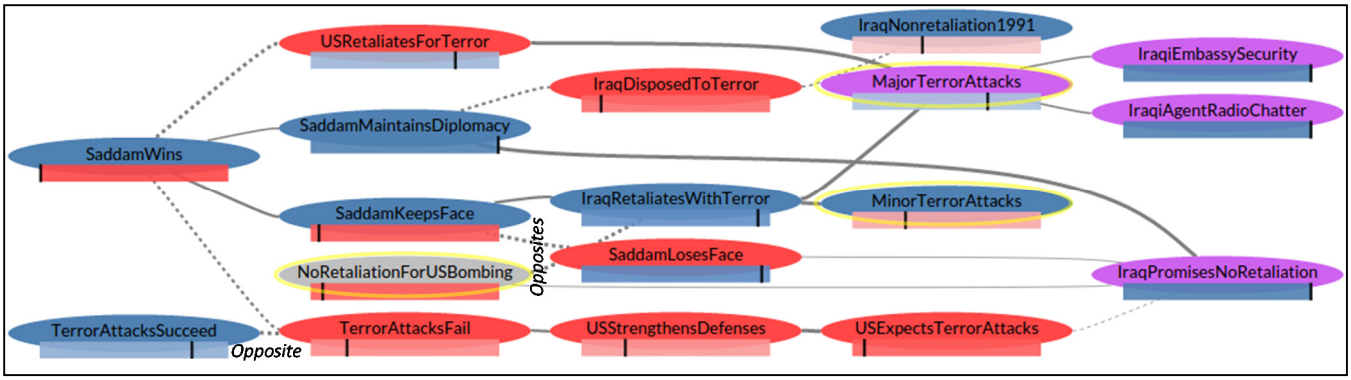


Figure 3: Statement nodes are connected by positive (solid grey line) and negative (dashed grey line) indication links of various strengths (per line thicknesses). Argument flow (from evidence to outcomes) is from right to left—e.g., SaddamWins is strongly indicated by SaddamKeepsFace. Outcome hypothesis nodes are circled in yellow. SaddamWins (hard finding false) captures Saddam's incentives to act or not. Belief bars' tick marks fall on a linear scale. Colors are explained in (Schrag et al., 2016a), also (Schrag et al., 2016b).

In developing the model in Figure 3, we identified some representation and reasoning shortcomings for which we are now implementing responsive capabilities (Schrag et al., 2016b). Relevant to our discussion here, TerrorAttacksFail (likewise TerrorAttacksSucceed) should be allowed to be true only when TerrorAttacks also is true.

We are working towards Logic nodes supporting any propositional expression using unary, binary, or higher arity operators[9]. When a Logic statement has a hard true finding[10], we refer to it as a Logic constraint, otherwise as a summarizing Logic statement.

We know that an attempted action can succeed or fail only if it occurs. By explicitly modeling (as Hypotheses) both the potential action results and adding a Logic constraint[11], we can force zero probability for every excluded truth value combination, improving the model. See Figure 4. The constraint node (left, in right model fragment) ensures that the model will believe in attack success/failure only when an attack actually occurs. Setting the hard true finding on this node turns the summarizing Logic statement (left, in the left fragment) into the Logic constraint—but also distorts the model's computed probabilities for the three Hypotheses. Presuming these probabilities have been deliberately engineered by the modeler, our framework must restore them. It does so by implementing (bottom fragment) a target belief (per the ConstraintTBC node) on one of the Hypotheses.

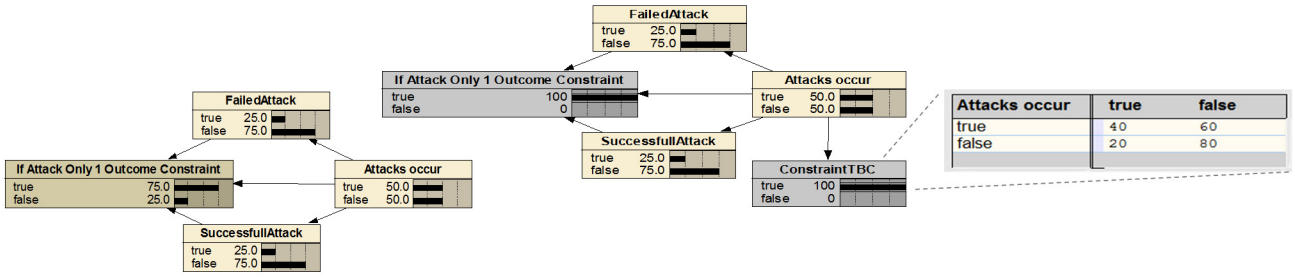


Figure 4: Logic constraints can help ensure sound reasoning.

[9] See, e.g., https://en.wikipedia.org/wiki/Truth_table.

[10] A likelihood finding could be used to implement a soft constraint.

[11] This constraint can be rendered (abbreviating statement names) as (or (and occur (xor succeed fail)) (and (not Occurs) (nor Succeeds Fails))) or more compactly via an if-then-else logic function (notated ite) as (ite Occurs (xor Succeeds Fails) (nor Succeeds Fails))—if an attack occurs, it either succeeds or fails, else it neither succeeds nor fails.

We implement a target belief either (depending on purpose) using a BN node like ConstraintTBC or (equivalently) via a likelihood finding on the subject BN node. The GUI does not ordinarily expose an auxiliary node like ConstraintTBC to a SME/analyst-class user.

This example is for illustration. We can implement this particular BN pattern without target beliefs. We also could implement absolute-strength IndicatedBy links as simple implication Logic constraints. However, this would not naturally accommodate one of these links' key properties—the ability to specify degree of belief in the link's upstream node when the downstream node is true—relevant because we can infer nothing about P given $P \Rightarrow Q$ and knowing Q to be true. It also demands two target belief specs that tend to compete. We are working to identify more Logic constraint patterns that can be implemented without target beliefs and to generalize specification of belief degree for any underdetermined entries in a summarizing Logic statement's CPT.

5. TARGET BELIEF PROCESSING

Ben Mrad et al. (2015) survey BN inference methods addressing fixed probability findings—our target beliefs. The most recent published results (Peng et al., 2010) address problems with no more than 15 nodes (all binary). Apparently, earlier approaches materialized full joint distributions—these authors anecdotally reported late-breaking results using a BN representation, with dramatically improved efficiency. Mrad et al. report related capabilities in the commercial BN tools Netica and BayesiaLab. Netica's "calibration" findings are concerned with comparing predictions to real data and could help identify where target beliefs were needed, however would do nothing to satisfy them. We have not experimented with BayesiaLab. While our performance results may similarly be construed as anecdotal—we have not systematically explored a relevant problem space—we have addressed a much larger problem. Our person risk assessment BN includes over 600 nodes and 26 target beliefs.

The basic scheme of our target belief processing approach is to interleave applications of Jeffrey's rule[12] with standard BN inference. Intuitively, each iteration—or "fitting step" (Zhang, 2008)—measures the difference between affected nodes' currently computed beliefs and specified target beliefs, makes changes to bring one or more nodes closer to target, and propagates these changes in BN inference. We continue iterating until a statistic over computed-vs.-target belief differences meets a desired criterion, or until reaching a limit on iterations, in which case we report failure. Just as for hard findings and likelihood findings, not all sets of target beliefs can be achieved simultaneously. In our intended incremental model development concept of operations (CONOPS), the framework's report that a latest-asserted target belief induces unsatisfiability should be taken as a signal that a modeling issue requires attention—much as would the similar report about a latest-asserted CPT.

We have implemented the following refinements to this basic scheme, improving performance.

1. Measure beliefs on a (modified) log odds scale.
2. Conservatively[13] apply Jeffrey's rule to all affected nodes in early iterations/fitting steps, then in late steps select for adjustment just the node with greatest difference between computed and target beliefs.
3. Save the work from previous target belief processing for a given model (e.g., under edit) to support fast incremental operation.

5.1 MODIFIED LOG ODDS BELIEF MEASUREMENT

Calculating the differences between beliefs measured on a scale in the log odds family, vs. on a linear scale, better reflects differences' actual impacts. We use the function depicted in Figure 5—a variation on log odds in which each factor of 2 less than even odds (valued at 0) loses one unit of distance that we refer to as a bit. So, for belief = 0.125 we calculate –2 bits.

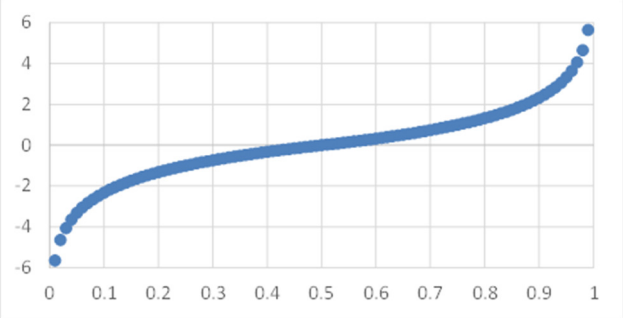


Figure 5: Belief transformation function (modified log odds) used in calculating computed-vs.-target belief differences

We express differences between beliefs in terms of such bits. So, difference(0.999, 0.87) = 7.02 bits and difference(0.87, 0.76) = 0.90 bits, whereas both pairs of untransformed beliefs (that is, (0.999, 0.87) and (0.87, 0.76)) have the same ratio, 1.14.[14] The transformation

[12] See (Jeffrey, 1983), as mentioned in section 1.

[13] See section 5.2.

[14] This difference metric is more conservative than the Kullback-Leibler distance or cross-entropy metric used in (Peng et al., 2010)'s I-divergence calculation. The absolute value of this function also has the advantage of being symmetric.

seems to inhibit oscillations among competing target beliefs.

5.2 MULTIPLE ADJUSTING IN ONE FITTING STEP

Moving all affected nodes all the way to their target beliefs in one fitting step is too aggressive in this model. We can get closer to a solution by adjusting more conservatively. We found that applying Jeffrey's rule to take affected variables {½ , 1/3, ¼, ...} of the way toward their target beliefs in successive fitting steps worked better than scaling calculated differences by any fixed proportion. This trick seems to be advantageous just for the first two or three fitting steps, after which single-node adjustments become more effective.

Incorporating both this refinement and the preceding one and running with a maximum belief difference of 0.275 bits for any node (yielding adequate model fidelity for our application), we complete target belief processing in 19 seconds (running inside a Linux virtual machine on a 2012-vintage Dell Precision M4800 Windows laptop).[15] That's not necessarily GUI-fast, but this is a larger model than many of our SME users may ever develop. Fitting steps took a little less than one second on average, with each step's processing dominated by the single call to BN inference.

These results remain practically anecdotal, as we have so far developed in our framework only this one large model including many target beliefs. Experience with different models may lead to more generally useful values for run-time parameters.

5.3 INCREMENTAL OPERATION

Under incremental operation, we execute only single-node fitting steps, as individual model edits usually have limited effect on overall target belief satisfaction. So far, we have experimented with incremental operation only for our person risk model.

Over two runs (with target beliefs processed in original input order vs. reversed):

- Average processing times per affected node were 2.1 and 2.3 seconds, respectively. Individual target beliefs processed in about 1.1 seconds or less about half the time. Figure 6 plots processing times for the first run, by affected node number, including a 4-node moving average.
- The least number of fitting steps was 0, the greatest 17 (taking from 0 to 8.7 seconds).
- Total run times were 54 and 59 seconds, respectively. So, batch (vs. incremental) processing can be advantageous, depending on CONOPS and use case.

[15] We found that tightening tolerance by a factor of 6.6 increased run time by a factor of 3.0.

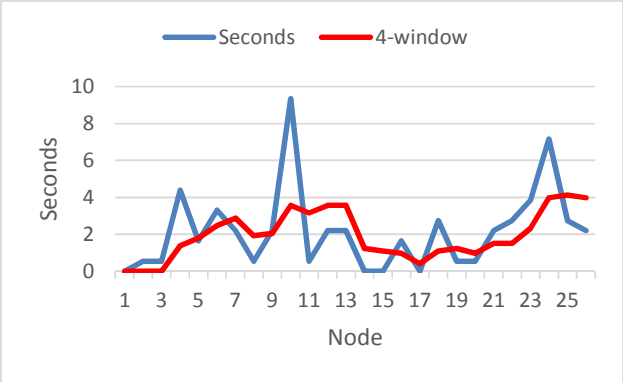


Figure 6: Run-time by affected node increment, with 4-node moving average window

6. CONCLUSION

Target beliefs have an important place in our SME-oriented modeling framework, where their processing is supported effectively by our methods described here. We might reduce or eliminate requirements for exogenous target beliefs by pushing SMEs towards arbitrary-precision link strengths (see Schrag et al. 2016b), but we are counting on target belief machinery to implement Logic constraints that make the SMEs' accessible modeling representation more expressive and versatile—ultimately more powerful. We expect target belief processing to be well within GUI response times for small models, including, per (Burns, 2015), the vast majority of intelligence analysis problems amenable to our argument mapping approach. We anticipate further work, especially to develop theory and practice for efficient implementation of different Logic constraint patterns.

Acknowledgements

We gratefully acknowledge the stimulating context of broader collaboration we have shared with other co-authors of (Schrag et al., 2016a; 2016b).

References

Ali Ben Mrad, Veronique Delcroix, Sylvain Piechowiak, Philip Leicester, and Mohamed Abid (2015), "An explication of uncertain evidence in Bayesian networks: likelihood evidence and probabilistic evidence," *Applied Intelligence,* published online 20 June 2015.

Kevin Burns (2015), "Bayesian HELP: Assisting Inferences in All-Source Intelligence," *Cognitive Assistance in Government,* Papers from the AAAI 2015 Fall Symposium, 7–13.

CIA Directorate of Intelligence, "A Tradecraft Primer: The Use of Argument Mapping," *Tradecraft Review* 3(1), Kent Center for Analytic Tradecraft, Sherman Kent School, 2006.

Richards J. Heuer, Jr. (2013), *Psychology of Intelligence Analysis,* Central Intelligence Agency Historical Document. https://www.cia.gov/library/center-for-the-study-of-intelligence/csi-publications/books-and-monographs/psychology-of-intelligence-analysis (Posted: Mar 16, 2007 01:52 PM. Last Updated: Jun 26, 2013 08:05 AM.)

R. Jeffrey (1983), *The Logic of Decision, 2nd Edition,* University of Chicago Press.

Christopher W. Karvetski, Kenneth C. Olson, Donald T. Gantz, and Glenn A. Cross (2013), "Structuring and analyzing competing hypotheses with Bayesian networks for intelligence analysis," *EURO J Decis Process* 1:205–231.

Yun Peng, Shenyong Zhang, Rong Pan (2010), "Bayesian Network Reasoning with Uncertain Evidences," *International Journal of Uncertainty, Fuzziness and Knowledge-Based Systems,* 18(5):539–564.

Robert Schrag, Edward Wright, Robert Kerr, and Bryan Ware (2014), "Processing Events in Probabilistic Risk Assessment," *9th International Conference on Semantic Technologies for Intelligence, Defense, and Security (STIDS).*

Robert Schrag, Joan McIntyre, Melonie Richey, Kathryn Laskey, Edward Wright, Robert Kerr, Robert Johnson, Bryan Ware, and Robert Hoffman (2016a), "Probabilistic Argument Maps for Intelligence Analysis: Completed Capabilities," *16th Workshop on Computational Models of Natural Argument*.

Robert Schrag, Edward Wright, Robert Kerr, Robert Johnson, Bryan Ware, Joan McIntyre, Melonie Richey, Kathryn Laskey, and Robert Hoffman (2016b), "Probabilistic Argument Maps for Intelligence Analysis: Capabilities Underway," *16th Workshop on Computational Models of Natural Argument*.

Edward Wright, Robert Schrag, Robert Kerr, and Bryan Ware (2015), "Automating the Construction of Indicator-Hypothesis Bayesian Networks from Qualitative Specifications," Haystax Technology technical report, https://labs.haystax.com/wp-content/uploads/2016/06/BMAW15-160303-update.pdf.

Shenyong Zhang, Yun Peng, and Xiaopu Wang (2008), "An Efficient Method for Probabilistic Knowledge Integration," In *Proceedings of The 20th IEEE International Conference on Tools with Artificial Intelligence (ICTAI),* November 3–5, vol 2. Dayton, pp 179–182).

Bayesian Models to Assess Risk of Corruption of Federal Management Units

Ricardo S. Carvalho[1] **and Rommel N. Carvalho**[1,2]
[1]Department of Research and Strategic Information at the Brazilian Office of the Comptroller General *
[2]Department of Computer Science at the University of Brasília †
{ricardo.carvalho,rommel.carvalho}@cgu.gov.br

Abstract

This paper presents a data mining project that generated Bayesian models to assess risk of corruption of federal management units. With thousands of extracted features related to corruptibility, the data were processed using techniques like correlation analysis and variance per class. We also compared two different discretization methods: Minimum Description Length Principle (MDLP) and Class-Attribute Contingency Coefficient (CACC). The feature selection process used Adaptive Lasso. To choose our final model we evaluated three different algorithms: Naïve Bayes, Tree Augmented Naïve Bayes, and Attribute Weighted Naïve Bayes. Finally, we analyzed the rules generated by the model in order to support knowledge discovery.

1 INTRODUCTION

Currently, it is known that corruption is a recurrent and primary subject on the Brazilian government agenda, fundamentally requiring its ostensive and efficient combat. Public corruption can be defined – supported by Brazilian Law no. 8,429, of June 1992[1] – as the act of misconduct or improper use of public office that leads to illicit enrichment, causing injury to the public treasury or infringing upon the principles of the public administration.

The Brazilian Office of the Comptroller General (CGU), an agency incorporated in the Presidency structure, has as one of its competences the role of assisting directly and immediately the President on matters and measures related to preventing and fighting corruption. Through activities of strategic information production, the Department of Research and Strategic Information (DIE) is the area responsible for investigating possible irregularities involving federal civil servants working in management units.

Nowadays, there are more than thirty thousand active federal management units[2], all subject to investigation. Due to this large number of units, most of the time DIE is limited to performing only investigations of those involved on large federal operations or recurrent complaints, often restricting its activities to cases triggered externally. Thus, it is important to have prioritization of activities based on risks of involvement in corruption so that DIE can act more effectively and proactively.

This work has two main objectives and contributions. The first is to build a Bayesian model to assess risk of corruption of federal management units. To this end, we seek to apply data mining techniques based on the state-of-the-art, along with a practical study of the information related to corruption. Therefore, we wish to contribute to CGU's activities in fighting corruption by building an useful model for their work priorization. Also, the step-by-step of this data mining project might be interesting for other practitioners, since it involves the combination of several different methods. We show how we applied correlation analysis and two discretization methods to process features, Adaptive Lasso for feature selection, and end up comparing three different algorithms to choose our final Bayesian model. Hence, this work gives contribution to practitioners while describing the application of data mining techniques with a practical objective and singular combination of techniques.

*SAS, Quadra 01, Bloco A, Edificio Darcy Ribeiro Brasilia, DF, Brazil
†Campus Darcy Ribeiro Brasília, DF, Brazil

[1]Brazilian Law no. 8.429, June 1992: http://www.planalto.gov.br/ccivil_03/leis/l8429.htm

[2]Management Units dataset: http://www.tesourotransparente.gov.br/ckan/dataset/siafi-relatorio-unidades-gestoras

The second objective is to achieve knowledge discovery in relation to information about corruptibility of federal management units, seeking to extract new rules in this domain. To this end, the information of management units available – as well as its direct and indirect relationships with the federal civil servants working there – are analyzed with the support of DIE experts in fighting corruption. After building our final model, we analyzed its derived rules. With this in mind, we wish to contribute to the enrichment of the experts' knowledge in fighting corruption.

In Section 2, we depict works most closely related to fighting corruption and how data mining has been used, while in Section 3 we give an overview of the information selected by DIE experts that will be used to build our models. Section 4 describes steps taken to pre-process data, such as correlation analysis, discretization, and also feature selection. In Section 5 we show how we used machine learning to build several models and Section 6 depicts our evaluation strategies. Section 7 discusses our deployment efforts related to the products of this work and we end this paper with a conclusion in Section 8.

2 RELATED WORK

In the last decade, observing current research areas, a topic closely related to risk of corruption is fraud detection. The main objective of fraud detection is to reveal trends of suspicious acts. For example, an emerging theme is to use data mining to detect financial fraud. A review of the academic literature of such application (Ngai et al., 2011) shows its successful use in detecting credit card fraud, money laundering, bankruptcy prediction, among others. This review also identifies common data mining techniques used in fraud detection, including Artificial Neural Networks, Decision Trees, Logistic Regression, and Naïve Bayes.

In this context, a recent survey on the subject of data mining-based fraud detection (Phua et al., 2010) displays a summary of published technical articles and a review on the topic. This survey, as well as other works (Kou et al., 2004), includes comments on similar applications. Also, an individual-oriented corruption analysis (Carvalho et al., 2014) was done building a corruption risk model for affiliated civil servants with algorithms like Random Forest and Bayesian Networks.

Regarding aspects of corruption, research related to public bidding and contracting processes has also been carried out, though not as widely as in fraud detection. The use of clustering and association rules to the problem of cartels in public bidding processes (Silva and Ralha, 2010) found results that corroborate the application of data mining in the prevention of corruption. Another paper (Balaniuk et al., 2012) shows the use of Naïve Bayes to evaluate the risk of corruption in public procurement. The authors applied natural logarithm to discretize attributes and based their assessment on the results of the conditional probabilities defined by experts.

In addition, a recent paper (Carvalho et al., 2013) presents the use of probabilistic ontologies to design and test a model that performs the fusion of information to detect possible fraud in bidding processes involving federal money in Brazil.

With respect to discretization algorithms, it has currently received a lot of focus as a pre-processing technique, mostly since many machine learning algorithms are known to produce better models by discretizing continuous attributes (Garcia et al., 2013). Two algorithms have received generally great performance, namely: CACC (Class-Attribute Contingency Coefficient) (Tsai et al., 2008) and MDLP (Minimum Description Length Principle) (Irani, 1993). In this work we compare the results of these algorithms after feature selection by creating models to allow us to choose the best results.

For feature selection, a recent review (Tang et al., 2014) shows several different widely used techniques, such as Adaptive Lasso (Zou, 2006). The Adaptive Lasso has basically two steps. First, an initial estimator is obtained, usually using Ridge Regression (Zou, 2006). Then a optimization problem with a weighted L1 penalty is carried out. The initial estimator generally puts more weight on the zero coefficients and less on nonzero ones to improve upon its predecessor: the Lasso (Zou, 2006). Compared to the Lasso, the adaptive Lasso has the advantage of the oracle property (Zou, 2006), resulting in a performance as well as if the true underlying model were given in advance. Compared to the SCAD and bridge methods (Tang et al., 2014), which also have the oracle property, the advantage of the adaptive Lasso is its computational efficiency.

3 DATA UNDERSTANDING

Seeking to analyze corruptibility of federal management units, various databases that DIE has access have been identified as useful for this work. For a better understanding of the data, the available information were divided into four dimensions, namely: Corruption; Employment; Sanctions; and Political.

Some of the information treated in this work are related to the federal civil servants that work in the management units. These information can give an idea of how much power a certain unit concentrates or how much influence the civil servants bring to the unit environment.

Due to the limited size of this paper, we present each dimension giving only an overview of the existing databases and relevant information identified by DIE experts regarding possible relationships with corruptibility.

3.1 CORRUPTION DIMENSION

CGU maintains the Federal Administration Registry of Expelled (CEAF)[3], which is a database with information that gathers expulsion penalties (expel, retirement abrogation, and dismissal) of federal civil servants since the year of 2003.

This database will be used to define management units that are corrupt, namely the positive class in our machine learning algorithms. The first paragraph of the Section 4 describes how this is done.

3.2 EMPLOYMENT DIMENSION

The employment dimension covers the information of management units regarding the federal civil servants that work there. It may be related to basic information such as office time and income, or even data that exposes the importance of the unit the servant is working – such as number of coordination roles or critic public offices like those that deal directly with public resources or financial assets.

Most of the information comes from the Human Resources Integrated System (SIAPE) of the Brazilian Federal Government[4].

For the employment dimension, the experts in fighting corruption of DIE selected 16 different information, that later can be transformed in 16 or more different features in the data preparation phase. Examples of these information are: mean, maximum, and minimum monthly income; number of coordination roles that deal with public contracts; number of roles for specific activities such as head of regional agency.

3.3 SANCTIONS DIMENSION

The sanctions dimension covers the information of management units that got sanctioned, due to practices of bad management of public money. We used sanctions in the Accounts Judged Irregular (CADIRREG) from the Federal Court of Accounts (TCU)[5], that judges the accounts of each management unit, deciding about its regularity according to Brazilian laws. Similarly, we used CGU's certificates of management irregularity[6].

Therefore, the experts in fighting corruption of DIE selected four different information, that later can be transformed in four or more different features in the data preparation phase. Examples of these information are: number of accounts judged irregular from TCU; and number of regularity certificates from CGU.

3.4 POLITICAL DIMENSION

The political dimension covers data of federal civil servants related to political activities, namely analyzing information of affiliation to political parties. By getting the affiliated servants of each management unit, we can measure how much each political party influences the units and if this will relate to corruption. The main database comes from Superior Electoral Court (TSE)[7].

Taking into account the knowledge of DIE experts, from the data provided by TSE we selected nine different information. Examples are: number of affiliations for a given political party and total number of affiliated servants in each management unit.

4 DATA PREPARATION

The data to be prepared are extracted for two classes, called "Corrupt" and "Non Corrupt". On one hand, "Corrupt" management units are those that throughout its history have had at least one civil servant who was expelled due specifically to corruption. In other words, units that had corrupt civil servants, which are those registered in CEAF whose legal basis for expulsion is consistent with our definition for corruption, as stated in Section 1.

On the other hand, to build the "Non Corrupt" group, we sampled a large group of management units and removed those considered "Corrupt" by definition, keeping the random sample proportion.

Thus, the dataset for non corrupt was created with a random sample of approximately 4,800 federal management units – amount approximately 8 times greater than the number of corrupt units.

[3]CEAF – Link: http://www.portaldatransparencia.gov.br/expulsoes/entrada

[4]Website for the Human Resources Integrated System (SIAPE) of the Brazilian Federal Government: http://www.siapenet.gov.br

[5]CADIRREG: http://contas.tcu.gov.br/cadirreg/CadirregConsultaNome

[6]CGU's audits reports: http://sistemas.cgu.gov.br/relats/relatorios.php

[7]TSE repositories: http://www.tse.jus.br/eleicoes/estatisticas/repositorio-de-dados-eleitorais

The data preparation phase includes feature selection and goes through the following steps, which will be described in the next sections:

- Data Cleaning and Feature Engineering: Adjusts the dataset;
- Preliminary Analysis: Treats variance zero per class and correlation;
- Data Separation: Segregates data for training and testing;
- Intermediary Analysis: Variance and correlation filtering;
- Feature Selection: Uses Adaptive Lasso;
- Discretization: Applies MDLP and CACC;

4.1 DATA CLEANING AND FEATURE ENGINEERING

Besides usual data cleaning activities – such as adjustment of inconsistencies, data conversion, and standardizing data types – the treatment of missing values was also conducted. For categorical variables we created a category "NA" representing the absence of values for a given variable. As for counting numerical variables, missing values represent the actual value of zero, so they were replaced by such value. In addition, other fields with missing values were treated individually. For example, date of cancellation of party affiliation, when affiliation still active, were replaced by a current date in order to create features for time of affiliation.

On feature engineering, first we created binarized features for all the categorical variables. Then, since some information can be registered more than once for a given management unit – for example, one can have several regularity certificates – we had to summarize the features for each unit. With only numerical features, a few of them were summarized by creating features with maximum, minimum, average, and total. For example, annual income was transformed into maximum annual income, minimum annual income, and mean annual income.

After this step, we had created 2,238 different features.

4.2 PRELIMINARY ANALYSIS

At first we removed features that had variance, within one of the classes, equal to zero, since with zero class-variance algorithms might bring estimates of coefficients that do not generalize (Hosmer et al., 2013). After calculating class-variance for each of the 2,238 features, 747 of them were removed – most of these being related to binarized categorical variables.

We also preliminarily addressed perfect pairwise correlation, which accounts for redundant information and may give biased estimates. Perfectly correlated features may have been added accidentally, or may have arisen after feature engineering.

Among the 1,495 variables analyzed, 96 – 48 pairs – returned perfect correlation. DIE experts chose which to eliminate in each pair.

4.3 DATA SEPARATION

At this point, our complete dataset had 688 corrupt units and 4,792 non corrupt units, with 1,447 features.

In this step we created two different datasets: Training Data (DT) and Testing Data (DTE). The first will be used through all data preparation and modeling, while the second will only be used as a final test after choosing the best final model.

To keep the original balance, DTE was created using a random sample of 20% of corrupts plus 20% of the non corrupts, and DT stayed with the remaining data, corresponding to 80% of the complete dataset.

4.4 INTERMEDIARY ANALYSIS

Similarly to the Preliminary Analysis, we again analyzed the class-variance. This resulted in removing 62 features with zero variance in one of the classes.

Nevertheless, in the intermediary analysis we did a different correlation analysis, following the well known hypothesis (Hall, 1999): "A good feature subset is one that contains features highly correlated with (predictive of) the class, yet uncorrelated with (not predictive of) each other".

Initially we calculated the correlation matrix of the remaining 1,376 features, also adding their correlation with the class column indicating corruptibility – 0 to non corrupt units and 1 to corrupt units. Then we filtered pairs of features with correlation equal or greater than 0.70 (absolute value) – number generally considered high correlation (Taylor, 1990). After that, the resulting matrix was sorted in descending order regarding the correlation of the features in relation to the class.

Thereafter, the rows of the matrix were traversed from the features with the largest correlation to the class. In each row, we kept the feature with the highest correlation with the class and removed the remaining features – from the dataset and the matrix – that had inter-correlation higher than 0.70 (absolute value).

With this algorithm we eliminated 468 features that had absolute correlation equal or greater than 0.70, thus remaining 910 features.

Such an approach was used to try to avoid the collinearity problem, mainly due to the fact that it is impossible to analyze all the possible combinations of feature groups, involved in this work. Thus, the correlation heuristic of each feature with its class – although not fully reflected in a model due to interactions between the features – serves as a technique to try to keep the theoretically most significant features – considering the correlation with class[8].

4.5 FEATURE SELECTION

To perform feature selection, each dataset passes through a regularized regression, specifically using Adaptive Lasso. For this purpose, we start by performing Ridge Regression with 10-fold cross-validation on the DT dataset. The estimates of the coefficients are used to construct an adaptive weights vector. With this vector introduced as the penalty factor, we implement Adaptive Lasso with 10-fold cross-validation. It is worth noticing that the Adaptive Lasso can force some of the coefficients to have estimates exactly equal to zero, thereby reducing the number of features.

After feature selection with Adaptive Lasso, we selected 144 features. The 10-fold cross-validation resulted in a AUC (Area Under the ROC Curve) (Bradley, 1997) of 0.85, considered satisfactory.

4.6 DISCRETIZATION

In recent years, discretization has received increasing research attention (Garcia et al., 2013). In the case of non-monotonic variables, the use of discretization techniques proves to be essential since it makes it possible to separate an original non-monotonic variable in various monotonous derived covariates (Tufféry, 2011). Also, when thinking about Bayesian models, some algorithms need all the features to be categorical, and discretization is a method of doing so.

In recent research (Garcia et al., 2013), two algorithms have received generally great performance, namely: MDLP (Minimum Description Length Principle) (Irani, 1993) and CACC(Class-Attribute Contingency Class) (Garcia et al., 2013). We compare these algorithms by later creating models for groups of features discretized with each method.

Accordingly, we have generated two different datasets from DT, one dataset for each discretization method used. The dataset discretized with MDLP algorithm returned 23 binary features, while CACC returned 66 – the reason these datasets have less features than the original is due to the fact that constant features were automatically removed.

[8]It may be useful to use different methods to analyze correlation in future work.

5 MODELING

In the modeling phase we started by creating models for each of the datasets discretized with MDLP and CACC. For this, we created Bayesian models using three different algorithms: Naïve Bayes (Lowd and Domingos, 2005), Tree Augmented Naïve Bayes (Zheng and Webb, 2011), and Attribute Weighted Naïve Bayes (Taheri et al., 2014).

This task was done using the R Package named caret[9]. We used 10-fold cross-validation to evaluate AUC and tried several different combinations of parameters for each of the three algorithms – from 20 to 60 combinations. For example, for Tree Augmented Naïve Bayes we used three score functions (loglik, bic, aic) each along side 20 different values for smoothing (from 0 to 19). After these models were built, caret selects the one with the combination of parameters that resulted in the best AUC value for each algorithm.

5.1 DISCRETIZATION SELECTION

The first step is to choose the most suitable discretization. With this in mind, for each discretized dataset we take the average results of AUC for the three algorithms used, again using 10-fold cross validation to try to estimate the out-of-sample results. The mean AUC outcomes are depicted in Table 1, along side the number of features each dataset has.

Table 1: Mean Results of Bayesian Models for each Discretized Dataset

Discretization	No. of features	**AUC**
MDLP	23	0.82
CACC	66	0.83

Although the results for the dataset with CACC discretization were slightly better, it is desirable to minimize the number of features considered in a model. Mainly models with less features tend to be more numerically stable and be adopted more easily. Also, a model with less features can avoid overfitting and increase its interpretability.

[9]R Package caret: `https://cran.r-project.org/web/packages/caret/index.html`

Therefore, we chose to select the features discretized with MDLP, since the respective model achieved results close to CACC but kept almost three times less features.

5.2 MODEL SELECTION

With the discretized dataset chosen, we now evaluate the Bayesian models built with the three algorithms: TAN (Tree Augmented Naïve Bayes) AW-NB (Attribute Weighted Naïve Bayes) and NB (Naïve Bayes). The AUC outcomes are showed in Table 2.

Table 2: Results of Bayesian Models for MDLP Dataset

Algorithm	**AUC**
TAN	0.8272
AW-NB	0.8207
NB	0.8244

Observing the results we chose the Bayesian model created with NB (Naïve Bayes) to be our final model, since it is more interpretable and simpler, while keeping practically the same results as the other two models.

6 EVALUATION

In the evaluation phase, we start by analyzing the results of the final model on the testing data separated on the beginning of this work. Finally, we analyzed the conditional probabilities of the features to extract useful knowledge regarding fighting corruption.

6.1 TESTING DATA

To ultimately validate our final model, we used the dataset separated in the data preparation phase for this purpose: the testing dataset (DTE). The first step here is to adjust DTE to have the same 23 final features selected from MDLP discretization.

Then, applying the final model on DTE we got AUC of approximately 0.76. Hence, we consider the results satisfactory. The reason being that the results are just a little below those obtained in the training dataset and are higher than 0.70, considered to be a threshold of good models.

6.2 KNOWLEDGE DISCOVERY

Observing the conditional probabilities of the final model, we extracted the rules it follows to define corruptibility for federal management units. This knowledge discovery aims to give a contribution to the activities of fighting corruption. Some of the main rules extracted that indicate an increase of risk of corruption are showed below.

- Accounts judged irregular by TCU;
- Responsibilities related to financial activities;
- Substitution public functions for controlling expenses;
- Number of requested civil servants allocated;
- Heading roles on regional agencies;
- Political party affiliations;
- Activities spread by multiple municipalities; and
- Number of public offices occupied by designation (without a selective process).

After discussing the main rules with DIE experts, they made a few comments in order to rationalize upon the knowledge discovered by the model.

- Accounts judged irregular by TCU are themselves by definition scenarios that involve inadequacies or irregularities;
- Responsibilities related to expenses and financial activities are critical, since they involve public resources and possible embezzlements;
- A management unit with several civil servants allocated by request might show a scenario of poor strength of the internal career;
- The heading roles related to regional management units usually have civil servants holding a relatively high amount of decision-making power with greater discretion, displaying a scenario of high propensity to corruption;
- Political party affiliations are related to greater political influence in decisions of public interest on the federal management units;
- Units with activities on many municipalities have to deal with decentralization problems; and
- Public offices employed by designation are occupied in the government due to nomination from discretionary authorities, not necessarily related to merit.

Therefore, by analyzing the rules together with the experts' comments, we see that the results have reasonable suitability in scenarios involving federal management units.

7 DEPLOYMENT

In the deployment phase, we created a Web application to allow managers at CGU to query management units and analyze their risk of corruption. With paths of grouped queries, managers can now view management units organized by their agencies. They are also able to perform ad-hoc queries, using as input unique identifiers of management units to obtain risk of corruption analysis for an individual unit or groups of them.

To deploy the predictive model to assess risk of corruption we simply implemented the calculation of Naïve Bayes with the conditional probabilities for the features selected on our final model. Using the output probabilities given by the model, we then discretized the results manually to only show risk categories, specifically: less than 0.20 as Very Low; equal or greater than 0.20 but less than 0.40 as Low; equal or greater than 0.40 but less than 0.60 as Medium; equal or greater than 0.60 but less than 0.80 as High; and equal or greater than 0.80 as Very High.

The Web application also generates pdf reports containing, for a given management unit: risk of corruption, average and maximum risk of corruption of the management units on the same agency. The application not only shows risk results, but also several other government data related to each management unit, allowing a general view of each unit.

With the application running, we started to present this work to all areas of CGU. Currently, several activities involving management units are being prioritized using our risk of corruption predictive model together with other information.

8 CONCLUSION

This paper described a data mining project that generated Bayesian models to assess risk of corruption of federal management units. We analyzed data from several government databases and, with the help of DIE experts, we developed thousands of important features. These variables were prepared and pre-processed removing those with zero class-variance and high inter-correlation.

Feature selection was done using Adaptive Lasso, which selected the 144 most relevant features. We compared two different discretization methods: CACC and MDLP. Bayesian models were built for datasets discretized with the two methods using the following algorithms: Naïve Bayes, Tree Augmented Naïve Bayes, and Attribute Weighted Naïve Bayes. To first choose the best discretization method we evaluated our results obtaining the average of the 10-fold cross-validation metrics performed per dataset. MDLP was chosen due to great results aligned with a considerable reduction of the number of features selected – from 144 to 23.

After choosing the dataset discretized with MDLP we evaluated the AUC for the three algorithms used on modeling. The results were very close, approximately 0.82. Therefore, we chose the model created with Naïve Bayes to be our final model, since it is more interpretable and simpler.

The dataset labeled Testing (DTE) separated on data preparation was then used to confirm the validity of the final model. DTE showed AUC of approximately 0.76.

Finally, the rules of the final model were extracted. With help from DIE experts, we derived knowledge for corruption fight activities. Rules generated and experts' comments were outlined to give an overview of the results.

The predictive model from this project was also deployed in a Web application, allowing managers from CGU to query and analyze federal management units regarding their risk of corruption. With the results of our model, CGU is already prioritizing corruption related activities to help maximize audits efficacy.

Therefore, this work contributed with an end-to-end data mining project overview, with application of several state-of-the-art techniques. We reinforced CGU's activities in fighting corruption by building an useful model to assess risk of corruption of federal management units. The knowledge discovered is also increasing the expertise of DIE analysts. With the Web application developed from this project, we help potentially save millions in public resources. Additionally, with risk assessment we encourage proactive audits, helping managers plan their work. To that end, we generate impact nationwide in fighting corruption.

Acknowledgements

The authors would like to thank the corruption fighting expert Victor Steytler for providing useful insights for the development of this work. Finally, the authors would like to thank CGU for providing the resources necessary to work in this research, as well as for allowing its publication.

References

R. Balaniuk, P. Bessiere, E. Mazer, and P. Cobbe. Risk based Government Audit Planning using Naïve Bayes Classifiers. *Advances in Knowledge-Based and Intelligent Information and Engineering Systems*, 2012.

Andrew P Bradley. The use of the area under the roc curve in the evaluation of machine learning algorithms. *Pattern recognition*, 30(7):1145–1159, 1997.

Ricardo Carvalho, Rommel Carvalho, Marcelo Ladeira, Fernando Monteiro, and Gilson Mendes. Using political party affiliation data to measure civil servants' risk of corruption. In *2014 Brazilian Conference on Intelligent Systems (BRACIS)*, pages 166–171. IEEE, 2014.

Rommel Carvalho, Shou Matsumoto, Kathryn B. Laskey, Paulo C. G. Costa, Marcelo Ladeira, and Lacio L. Santos. Probabilistic ontology and knowledge fusion for procurement fraud detection in brazil. In *Uncertainty Reasoning for the Semantic Web II*, pages 19–40. Springer, 2013.

S. Garcia, J. Luengo, J. A. Sáez, V. Lopez, and F. Herrera. A survey of discretization techniques: taxonomy and empirical analysis in supervised learning. *Knowledge and Data Engineering, IEEE Transactions on*, 25 (4):734–750, 2013.

Mark A. Hall. *Correlation-based feature selection for machine learning*. PhD thesis, The University of Waikato, 1999. URL `https://www.lri.fr/~pierres/donn%E9es/save/these/articles/lpr-queue/hall99correlationbased.pdf`.

David W Hosmer, Stanley Lemeshow, and Rodney X Sturdivant. *Applied logistic regression*, volume 398. John Wiley & Sons, 2013.

Keki B Irani. Multi-interval discretization of continuous-valued attributes for classification learning. 1993.

Yufeng Kou, Chang-Tien Lu, Sirirat Sirwongwattana, and Yo-Ping Huang. Survey of fraud detection techniques. In *Networking, sensing and control, 2004 IEEE international conference on*, volume 2, pages 749–754. IEEE, 2004.

Daniel Lowd and Pedro Domingos. Naive bayes models for probability estimation. In *Proceedings of the 22nd international conference on Machine learning*, pages 529–536. ACM, 2005.

EWT Ngai, Yong Hu, YH Wong, Yijun Chen, and Xin Sun. The application of data mining techniques in financial fraud detection: A classification framework and an academic review of literature. *Decision Support Systems*, 50(3):559–569, 2011.

Clifton Phua, Vincent Lee, Kate Smith, and Ross Gayler. A comprehensive survey of data mining-based fraud detection research. *arXiv preprint arXiv:1009.6119*, 2010. URL `http://arxiv.org/abs/1009.6119`.

Carlos Vinícius Silva and Célia Ralha. Utilização de Técnicas de Mineração de Dados como Auxílio na Detecção de Cartéis em Licitações. In *WCGE - II Workshop de Computação Aplicada em Governo Eletrônico*, 2010.

Sona Taheri, John Yearwood, Musa Mammadov, and Sattar Seifollahi. Attribute weighted naive bayes classifier using a local optimization. *Neural Computing and Applications*, 24(5):995–1002, 2014.

Jiliang Tang, Salem Alelyani, and Huan Liu. Feature selection for classification: A review. *Data Classification: Algorithms and Applications*, page 37, 2014.

Richard Taylor. Interpretation of the correlation coefficient: a basic review. *Journal of diagnostic medical sonography*, 6(1):35–39, 1990.

Cheng-Jung Tsai, Chien-I Lee, and Wei-Pang Yang. A discretization algorithm based on class-attribute contingency coefficient. *Information Sciences*, 178(3): 714–731, 2008.

Stéphane Tufféry. *Data mining and statistics for decision making*. John Wiley & Sons, 2011.

Fei Zheng and Geoffrey I Webb. Tree augmented naive bayes. In *Encyclopedia of Machine Learning*, pages 990–991. Springer, 2011.

Hui Zou. The adaptive lasso and its oracle properties. *Journal of the American statistical association*, 101 (476):1418–1429, 2006.

The Efficacy of the POMDP-RTI Approach for Early Reading Intervention

Umit Tokac
Educational Psychology and
Learning Systems
Florida State University
Tallahassee, FL 32306
ut08@my.fsu.edu

Russell G. Almond
Educational Psychology and
Learning Systems
Florida State University
Tallahassee, FL 32306
ralmond@fsu.edu

Abstract

A POMDP is a tool for planning: selecting a policy that will lead to an optimal outcome. Response to intervention (RTI) is an approach to instruction, where teachers craft individual plans for students based on the results of progress monitoring tests. Current practice assigns students into tiers of instruction at each time point based on cut scores on the most recent test. This paper explores whether a tier assignment policy determined by a POMDP model in a RTI setting offer advantages over the current practice. Simulated data sets were used to compare the two approaches; the model had a single latent reading construct and two observed reading measures: Phoneme Segmentation Fluency (PSF) for phonological awareness and Nonsense Word Fluency (NWF) for phonological decoding. The two simulation studies compared how the students were placed into instructional groups using the two approaches, POMDP-RTI and RTI. This paper explored the efficacy of using a POMDP to select and apply appropriate instruction.

1. INTRODUCTION

Statistics gathered by local school districts reflect that roughly 30% of their first-grade students read below grade level standards (Matthews, 2015). Moreover, Landerl and Wimmer (2009) reported that 70% of struggling readers in first grade continued to struggle in eight grade when no intervention was provided. Mastropieri, Scruggs, and Graetz (2003) argued that reading is the main problem for most students with learning disabilities.

Torgesen (2004) asserts that reading consists of five components: phonological awareness, phonological decoding, fluency, vocabulary, and reading comprehension. According to the Simple View Theory of Reading Development (Gough & Tunmer, 1986) for children at young ages, mastery of the first two components, phonological decoding and phonological awareness, generate the remaining three reading components: fluency, vocabulary, and reading comprehension. A lack of either phonological decoding or phonological awareness affects the other components and causes reading difficulties. Because the development of reading skills is critical, instructors should identify children with reading difficulties and provide additional instructional support (Catts, Hogan & Fey, 2003).

Response to intervention (RTI) is an educational framework designed to identify students with difficulties in reading and math, and intervene as early as possible by providing more intensive instruction for students who need it. The RTI approach divides instruction into Tiers; each tier includes different intervention or instruction. The RTI process starts with screening tests which monitor general knowledge and skills of all students in the class. The screening tests are administered on multiple occasions during a school year. The screening test results provide teachers with a rough estimate of each student's proficiency that guides the assignment of students into appropriate tiers of instruction. RTI has produced good results in both research and operational settings, and hence is considered to be one of the evidence-based

practices for improving reading and preventing learning disabilities (Greenwood et al., 2011).

Ideally, the placement into Tiers of students in an RTI program would be based on their unobservable true proficiency. As this is unobservable, the placement decision is instead made basis of the estimates of proficiency from screening tests. Often in current practice this is implemented through a cut score on the most recent screening test. Naturally, a certain amount of measurement error causes some students to be placed incorrectly. Considering the entire (both students' previous screen-tests results and changes in instruction) history in account should improve the proficiency estimates performance. Almond (2007) suggested that this could be done using a partially observed Markov decision process (POMDP) — partially observed, because the true student proficiency is latent; a decision process, because the instructors decide what instruction or intervention to use between measurement occasions.

A POMDP is a probabilistic and sequential model. A POMDP can be in one of a number of distinct states at any point in time, and its state changes over time in response to events (Boutilier, Dean & Hanks, 1999). One noteworthy difference between a RTI approach and a POMDP model is that most RTI approaches use only the latest test results to identify students' proficiencies and assign them to appropriate tier (Nese et al., 2010). We call the approach the current-time only-RTI model. On the other hand, a POMDP-RTI model is the combination of a periodically applied screening test, and the RTI into a POMDP model. Additionally, a POMDP considers the students' entire histories (both actions and test scores) when determining appropriate interventions at in order to identify their current abilities and forecast their future abilities under competing policies. Therefore, a POMDP-RTI model should perform better than current-time only-RTI model.

To test the last assertion, this paper compares the POMDP-RTI model with the current-time only-RTI, evaluating the predictive accuracy of each model, the quality of the instructional plans produced and the reading levels achieved at the end of the year. It does this through simulation studies based on numbers obtained from fitting the POMDP model to a group of kindergarten students in an earlier RTI study (Al Otaiba, Connor, Folsom, Greulich, Meadows, & Li, 2011).

2. METHOD

Two simulated datasets were used in order to address how properly students are assigned to each tier based on their latent reading score in the POMDP-RTI model compared based on their observed score in the current-time only-RTI model. The initial value of the parameters were based on a longitudinal Florida Center for Reading Research (FCRR) study of reading proficiency (Al Otaiba et al, 2011) and data sets were simulated based on the Almond (2007) model in order to produce realistic data for answering the research question posed above. The parameters of the simulation were chosen so that the distribution of scores on the screening test were similar to those of the Al Otaiba et al. study at both the initial and final measurement period.

2.1 THE POMDP-RTI FRAMEWORK

Almond (2007) describes a general mapping of a POMDP into an educational setting. It is assumed that the student's proficiency is measured at a number of occasions. The latent proficiencies of the students is the hidden layer of the POMDP model. The actual test scores are the observable outcomes, and the instructional options for the teacher between measurement occasions are the action space. The utility is assumed to be an increasing function of the latent proficiency variable at the last measurement occasion; thus, it is finite time horizon model.

Figure 1 show a realization of an RTI program in this framework. The nodes marked R represent the latent student proficiency as it evolves over time. At each time slice, there is generally some kind of measurement of student progress represented by the observable outcomes, Phoneme Segmentation Fluency (PSF) for phonological awareness and Nonsense Word Fluency (NWF) for phonological decoding. Tiers are instructional tasks chosen by the instructor and applied during time slices. Note that in an RTI implementation, Tier 1 refers to whole class instruction given to all students, while Tier 2 is small group supplemental instruction generally given only to the students most at risk. Students in Tier 2 are given the Tier 1 instruction as well.

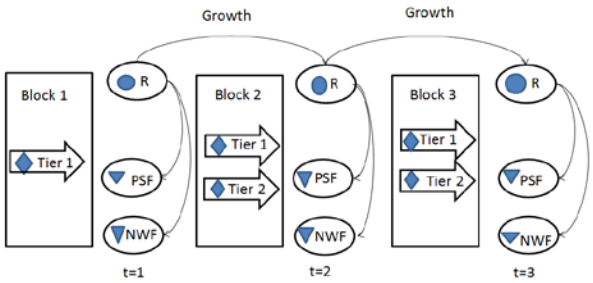


Figure 1: The POMDP-RTI model

The Figure 1 was designed based on evidence-centered assessment design (ECD; Mislevy, Steinberg, & Almond, 2003) we call this an *evidence model*. In general, both the

proficiency variables at Measurement Occasion m, R_m, and the observable multivariate outcome variables are PSF_m and NWF_m on that occasion. Extending the ECD terminology, Almond (2007) calls the model for the R_m's, the *proficiency growth* model. Following the normal logic of POMDPs this is expressed with two parts: the first is the initial proficiency model, which gives the population distribution for proficiency at the first measurement occasion. The second is an action, which gives a probability distribution for change in proficiency over time that depends on the instructional activity chosen between measurement occasions.

There are two notable differences between the POMDP models used in this application and those commonly seen in the literature. First, the models have a fixed and finite time horizon, with the reward occurring only at the last time step (although the actions at each step have a cost which is subtracted from the reward). This removes the need for the usual discounting of future rewards. The second is that the Markov process in non-stationary (it is hoped that the student's abilities will improve over time). This produces a potential identifiability issue, as growth is difficult to distinguish between difficulty shifts in the measurement instruments (Almond, Tokac & Al Otaiba, 2012). Assuming that the screening tests have all be equated, hence are on the same scale, takes care of the identification issue. An alternative approach would be to subtract the expected growth from the model, making the latent proficiency variable represent deviations from the expected growth model (Almond, et al., 2014).

2.1.1 Proficiency Growth Model

The model from which the data was simulated was a unidimensional model of reading with a single latent, continuous variable: R_{nm}, the reading ability of individual n on measurement occasion m. In this case, N was 300 students and M represented the three equally spaced time points, t_1, t_2, t_3. (RTI screening tests are typically given 3 times per year.)

This study assumed that a teacher provided general instruction to all the students until the first time point, t_1, and that the initial ability distribution was normal, $R_0 \sim N(0,1)$. As this is a purely latent variable, the scale and location is arbitrary. Fixing the initial population to have a standard normal distribution establishes the scale.

After analyzing the results of assessments administered at t_1, the teacher delivered additional and more intensive instruction to students who were assigned to *Tier 2*, but delivered only general instruction to students in *Tier 1*. The tier to which student n is assigned at time m is represented by $a(n,m)$. The growth rate for the students is assumed to depend on the tier assignment. Thus, for measurement occasion $m > 1$,

$$R_{nm} = R_{n(m-1)} + \gamma_{a(n,m)} \Delta T_m + \eta_{nm}, \qquad (1)$$

$$\text{where} \quad \eta_{nm} \sim N(0, \sigma_{a(n,m)}\sqrt{\Delta T_m}),$$

and where ΔT_m represents the elapsed time period between measurement occasions m and m-1 for *Tier 1 and Tier 2*. In this study, each school year was equal to 1, and ΔT_m was fixed and equal to 1/M (e.g. M = 3, so $\Delta T_m = \frac{1}{3}$). The parameter $\gamma_{a(n,m)}$ is a tier-specific growth rate and it was fixed and had two different initial values for each tier. We set $\gamma_{am} = 0.9$ for *Tier 1*, and $\gamma_{am} = 1.2$ for *Tier 2*. The residual standard deviation, $\sigma_{a(n,m)}\sqrt{\Delta T_m}$, depends on both a tier-specific rate, $\sigma_{a(n,m)}$, and the length of time, ΔT_m, between measurements (thus, growth is occurring via a non-stationary Brownian motion process). The standard deviation of the growth per unit time, $\sigma_{a(n,m)}$, was fixed to 1 for both tiers.

2.1.2 Evidence Model

The evidence model involved two independent regressions, one for each observed variable i. These two observable variables were chosen because they are critical reading components for later reading performance in the first two years of elementary school (Rock, 2007). Let Y_{nmi} be the observation for individual n at measurement occasion m on observed variable i of the proficiency variables, then:

$$R_{n0} \sim \mathrm{N}(0,1)$$

$$Y_{nmi} = a_i + b_i R_{nm} + \varepsilon_{nmi}, \qquad (2)$$

$$\varepsilon_{nmi} \sim N(0, \omega_i).$$

The reliability of the instruments can be used to determine b and ω. The reliability of an observed variable i at any time point was represented as r_i. In classical test theory, the reliability is the squared correlation coefficient between the true score and the observed score of the student. This definition translates into an equation as

$$r_i = 1 - (Var_n(\epsilon_{nmi}) / Var_n(Y_{nmi}))$$

where $Var_n(.)$ indicates that the variance comes from individuals (where measurement occasion and instrument are considered as constant). Then

$$b_i = \sigma_{Y_i}/\sigma_{R_i} * \sqrt{r^2} \quad \text{and}$$

$$\omega_i = \sigma_{Y_i} * \sqrt{1 - r^2}$$

In order to make $r_i = .45$ at each time point, t_m, for the measurement of each skill on observed variable i, $b_i = .98$ and $\omega_i = .65$ was used at t_m. These numbers are comparable to reading measures commonly used with 1st grade students. At this point, the model is very close to

the model described in Almond, Tokac and Al Otaiba (2012), except that the previous work assumed all students were in the same Tier. Appropriate values for *a* and *b* depend on the scale of the instruments chosen. The values used in the simulation were chosen so that the mean and standard deviation of the simulated data matched the data set from Al Otaiba et al. (2011) at the first and last time points.

2.1.3 Decision Rules

The key research question compares the performance of the system under two different policies. The first is a fixed decision rule implicit in the current-time RTI policy: Students who are below a cut-score on either of the two screening tests are placed into Tier 2 instruction. The second policy is the optimal policy found by solving the POMDP. Implementing this policy requires an explicit specification of the utility function and the cost function for the instructional options.

Many RTI implementations used the reference score (general class median score or some other percentile rank) as a cut score for assigning each student to either the Tier 1 or Tier 2 group. The simulated model used different Tier 2 for each of the two screening tests (NWF and PSF) giving four possible Tier assignments. For instance, if a student's score on the NWF test is lower than the cut score for NWF but higher for PSF, the student was assigned to Tier 2 for NWF and Tier 1 for PSF. (This differs slightly from the common practice which would put students who fail to meet the cut on either measure into a single Tier 2.)

The POMDP forecasts expected learning under each possible outcome and assigns students to tiers in a way that balances the expected learning gains with the cost of instruction. The utility function is the expected gain at the last time point and the cost function is the sum of costs of applied instruction at each state. The benefit is always higher for Tier 2, as is the cost. However, the cost exceeds the utility of the benefit for some regions of the distribution because the utility is nonlinear, while for other regions it does not.

The contact hours with the instructor drive the cost of each block. Cost is high for more intensive instruction in Tier 2, and, without loss of generality, it is zero for Tier 1, as all students receive Tier 1 instruction. The cost function consists of three components: the frequency with which the group meets, f_a, the duration of the meeting time, d_a, and size of the group, g_a (Almond & Tokac, 2014). Then

$$c(a) = k f_a d_a / g_a , \qquad (3)$$

represents the model cost of taking action or activity *a* in state *s,* where *k* is a constant used to put the cost function on the same scale as the utility function. In this study, the cost value was fixed at c(*Tier 2)* = 0.1 and c(*Tier 1)* = 0.

The utility function is

$$u(R_M) = logit^{-1}(\alpha(R_M - \beta)). \qquad (4)$$

In this equation α and β are fixed parameters; β is a proficiency target, which is on the scale of the internal latent variable R_M. Specifically β = 0.5 for *Tier 1* and β= 0.1 for *Tier 2*. Also, α is a slope parameter, and α= 0.8 for both *Tier 1* and *Tier 2*. High values of α favor bringing students near proficiency standards above the proficiency target β, while low values of α give more weight to enriching students at the high end of the scale and providing remediation at the low end of the scale (Almond & Tokac, 2014). (Almond & Tokac alternatively recommend using a probit function in place of a logit, so that α becomes effectively a standard deviation; however, the as the shape of the logit and probit curves are so similar, we expect the results using a probit curve would be similar as well.)

In this case, the total reward is $u(R_M) - c(a(s,2)) - c(a(s,3))$. The difference between the utility function and the cost function is the total reward for getting the student to proficiency level Tier 1 using instruction *a(s,2)* and *a(s,3)* between measurements 1 and 2, and 2 and 3. The reward is the basis for the assignment of each student to Tier 1 or Tier 2. The POMDP model forecasts the expected reward, and balances that with cost during each period.

2.2 SIMULATION DESIGN

The initial value of the simulated data student distribution at time 0 was based on the FCRR data set (Al Otaiba, 2007). In the FCRR data, the correlation between NWF and PSF was .65. The simulation generated latent proficiency variables for each simulee, and simulated scores on the reading scores on the NWF and PSF test administered at t_1, t_2 and t_3 in the model. At each time point, the correlation coefficient between NWF and PSF was around 0.65 and the same growth and measurement error residuals were used for both the POMDP-RTI and current-time only-RTI models.

The proficiency growth model and evidence model parameters were estimated from the simulated data through Markov Chain Monte Carlo (MCMC) simulation using JAGS (Plummer, 2003). Four independent Markov chains with random starting positions were used with 500000 iterations. This is consistent with standard practice (Gelman, Carlin, Stern & Rubin, 2004; Neal,

2010). Tokac (2016) describes tests done for convergence and parameter recovery with this model.

3. RESULTS

Data were simulated for students under two different policies, (1) current-time only-RTI policy where students are assigned to Tier 1 or Tier 2 based on a cuts scores on the PSF and NWF tests at the most recent time point, and (2) a POMDP-RTI policy where each student is assigned to the tier that maximizes the expected utility for that student. This resulted in two different simulated series: $R^{\check{}}_{nm}$ was the true reading ability under the current-time only cut score policy and $R^{\hat{}}_{nm}$ was the true reading ability under the POMDP-RTI policy. Note that the two simulations used the same residuals in equation (1) (growth residual η_{nm}) and equation (2) (measurement error ε_{nmi}). Thus, they differed only by the value of the growth rate parameter, $\gamma_{a(n,m)}$, used in equation (1).

Table 3: Comparison of the number of PSF and NWF scores between tiers categorized by cut scores or POMDP estimates

Method	Tier	PSF_{t2}	NWF_{t2}	PSF_{t3}	NWF_{t3}
POMDP	Tier 1	150	149	181	181
	Tier 2	150	151	119	119
Cut Score	Tier 1	150	149	150	150
	Tier 2	150	151	150	150

Table 3 shows the pattern of Tier assignment under the two models. At the second time point, the two policies behave roughly the same assigning the lowest performing 50% of students to Tier 2. However, at the third time point, substantially fewer students are assigned to Tier 2 under the POMDP-RTI policy. This might be a result of better placement policies, or simply that the Tier 2 support is less needed in the latter part of the school year.

Table 4 breaks down the differences between the two policies at time point 3. Recall that the students were classified into Tiers independently based on the PSF and NWF measures, resulting effectively in four different classifications: 1-1 (both in Tier 1), 1-2, 2-1 (mixed), and 2-2 (both Tier 2). Table 4 shows the number of students who were classified into one of the four groups who were classified into a different group by the other policy. Slightly over half (151) students were assigned different instruction under the different policies.

Table 4: Comparison of POMDP-RTI and Current-Time only-RTI models

	Number of Non-Matching Students	
	Time 3	
Tiers	POMDP - RTI	Current - Time RTI
1-1	49	20
1-2	38	36
2-1	42	40
2-2	22	55

Thus, there is a fair bit of difference in the placement, but which placement is better? As this is a simulation student, the true abilities are known it should be possible to determine an ideal placement based on the known simulated abilities. However, the abilities, $R^{\check{}}_{nm}$ and $R^{\hat{}}_{nm}$, are different in the two branches of the assessment (because a different policy was actually employed). Therefore, the ideal placements will be different under each policy.

In determining the ideal placement, the two mixed assignments, 1-2 and 2-1, were combined into a single mixed tier. Cut scores on the latent ability variable were calculated based on the utilities in equations (3) and (4) and a single growth step after the last measurement: the students with abilities higher than 0.1 should be placed into Tier 1, those lower than -0.4 into Tier 2 and students in between into the Mixed Tier. Both policies used the same cut points for determining the ideal placement, but because the abilities were different, the actual ideal placement could be different for the two students under the same policy at Time 3.

Table 5 presents the number of students placed in each tier under the actual and ideal placements under both policies. It also presents a measure of agreement which is the number of students assigned to that tier in the ideal placement that were actually assigned to the Tier. The POMDP-RTI does well under that metric, with all of the students who should be placed into Tier 1 or 2 correctly placed in that tier. This policy only had problems with the mixed tier, with 35% of the students being incorrectly placed in Tier 1 or Tier 2.

The current-time only-RTI policy did not fare as well. First, note that under the ideal placement for this policy fewer students would be in the high-performing Tier 1 group. This is likely due to incorrect assignment at Time 2. Next, note that agreement rates are lower. So the POMDP-RTI model did better on two important metrics.

To summarize the agreement numbers, we used Goodman and Kruskall's lambda (Almond, Mislevy, Steinberg, Yan, and Williamson, 2015). Usually, this adjusts the raw agreement rate by subtracting out the agreement with a classifier which simply classifies everybody at the modal category (which would be the mixed tier for both policies). However, Tier 1 has a special meaning in the context of RTI; Tier 1 is the normal whole-class instruction that is given regardless of the test score.

Table 5. Agreement between ideal and actual placement under POMDP-RTI.

Ideal Placement	POMDP-RTI Placement			
	Tier 1	Mix Tier	Tier 2	Total
Tier 1	118	0	0	118
Mix Tier	18	90	30	138
Tier 2	0	0	44	44
Total	136	90	74	300

Table 6. Agreement between ideal and actual placement under current-time only RTI.

Ideal Placement	Current-Time only-RTI Placement			
	Tier 1	Mix Tier	Tier 2	Total
Tier 1	72	17	0	89
Mix Tier	35	58	50	143
Tier 2	0	11	57	68
Total	107	86	107	300

Therefore, by using Tier 1 as the baseline in lambda, the result is a statistic that describes how much better the RTI is performing than undifferentiated whole class instruction. Let k_i be the number of students correctly classified into Tier i, and let k_{Tier1} be the number of students who should ideally be assigned to Tier 1. Then

$$\lambda = \frac{\sum_i k_i - k_{Tier1}}{N - k_{Tier1}}$$

Like a correlation coefficient, the value of lambda ranges between -1 and 1, with 0 representing a classifier which does no better than simply assigning everybody to the model category. If it is 1, it means that the policy did a perfect job of assigning students to the ideal tier. Using the data in Table 5, $\lambda = 0.74$ for POMDP-RTI, $\lambda = 0.51$ for Current-time only RTI. So RTI does better than undifferentiated instruction, but the POMDP-RTI policy also does better than the current-time only-RTI.

4. CONCLUSION

As expected, a policy produced by a POMDP (which is designed to produce optimal policies) performed better than current-time only cut-score policy current used in many RTI implementations. In particular, the POMDP-RTI had a better agreement with the ideal placement ($\lambda = 0.74$) than the current-time only model did ($\lambda = 0.51$). The likely reason for the better performance is that the POMDP model is better able to use the entire student record, both the history of assessments and instruction and multiple tests taken at the same time to build a more accurate estimate of student proficiency, although some may have been influenced by the use of the same utility model used in the POMDP to define ideal placement.

The cut-score approach currently in common use does have one clear advantage over the POMDP model: it is simpler to implement and explain. However, if the POMDP recommendations were integrated into an electronic gradebook, it might be better received by teachers. However, while teachers may not feel the need for the POMDP software to address the Tier 1/Tier 2 placement, there is another aspect of the RTI framework which was not addressed in this study. During Tier 2, students receive regular progress monitoring assessments, and the teacher is supposed to be making fine-grained adjustments if the student is not responding to the intervention (hence the name response-to-intervention). In particular, the teachers can adjust the intensity of the intervention (equation 3) adding more time on task if needed, or using less support if the teacher is appearing to do well. This is a target of opportunity for the POMDP model, as teachers have responded favorability to the idea of computer support to help them with tracking and intervention adjustment for Tier 2 students.[1] The present work shows that POMDPs are a promising approach to this problem.

Another limitation of the current work is that it assumes all students grow at the same rate under each of the instructional conditions (e.g., given the tier placement). In practice, many studies looking at RTI have found that students grow at different rates, with a low growth rate often corresponding to low initial ability.[2] While this adds complexity to the model, we think that the POMDP framework will help educators make optimal policy decisions with this additional information.

Acknowledgements

We would like to thank the Florida Center for Reading Research for allowing us access to the data used in this paper. The data were originally collected as part of a larger National Institute of Child Health and Human Development Early Child Care Research Network study.

References

Almond, R. G. (2007). Cognitive modeling to represent growth (learning) using Markov decision processes. *Technology, Instruction, Cognition and Learning (TICL), 5*, 313-324. Retrieved

[1] Joe Nese, U. Oregon, private communication. May 16, 2016.

[2] Young-Suk Kim, Florida State University. Private communication. March 31, 2016.

fromhttp://www.oldcitypublishing.com/TICL/TICL.html

Almond, R. G. (2011). *Estimating Parameters of Periodic Assessment Models* (Repot No. RM-11-06). Educational Testing Service. Retrieved from http://www.ets.org/research/policy_research_reports/rm-11-06.pdf

Almond, R. G., Mislevy, R. J., Steinberg, L. S., Yan, D., & Williamson, D. M. (2015). *Bayesian Networks in Educational Assessment*. Springer.

Almond, R., Goldin, I., Guo, Y., & Wang, N. (2014). Vertical and Stationary Scales for Progress Maps. In J Stamper, Z Pardoz, M Mavrikis, & B. M. McLaren (Eds.), *Proceedings of the 7th International Conference on Educational Data Mining, London, England*. Society for Educational Data Mining. 169—176. Retrieved from http://educationaldatamining.org/EDM2014/uploads/procs2014/long%20papers/169_EDM-2014-Full.pdf

Almond, G. R., Tokac, U., & Al Otaiba, S. (2012). Using POMDPs to Forecast Kindergarten Students' Reading Comprehension. In Agosta, J. M., Nicholson, A., & Flores, M. J. (Eds.), *The 9th Bayesian Modeling Application Workshop at UAI 2012*. Catalina Island, CA. Retrieved from http://www.abnms.org/uai2012-apps-workshop/papers/AlmondEtal.pdf

Almond, R. G., & Tokac, U. (2014, November). *Using Decision Theory to Allocate Educational Resources.* Paper presented at Annual Meeting, Florida Educational Research Association, Cocoa Beach, FL.

Almond, R. G., Yan, D., & Hemat, L. A. (2008). Parameter Recovery Studies with a Diagnostic Bayesian Network Model. *Behaviormetrika, 35*(2), 159-185.

Al Otaiba, S., Folsom, J. S., Schatschneider, C., Wanzek, J., Greulich, L., Meadows, J., & Li, Z. (2011). Predicting first grade reading performance from kindergarten response to instruction. *Exceptional Children*, *77(4),* 453-470.

Boutilier, C., Dean, T., & Hanks, S. (1999). Decision-theoretic planning: Structural assumptions and computational leverage. *Journal of Artificial Intelligence Research, 11*, 1-94. Available from citeseer.ist.psu.edu/boutilier99decisiontheoretic.html

Catts, H. W., Hogan, T. P. E., & Fey, M. (2003). Subgrouping poor readers on the basis of individual differences in reading-related abilities. *Journal of Learning Disabilities, 36*, 151–164.

Gelman, A., Carlin, J. B., Stern, H. S., & Rubin, D. B. (2004). *Bayesian Data Analysis*. Boca Raton, FL: Chapman and Hall.

Gough, P. B., & Tunmer, W. E. (1986). Decoding, reading, and reading disability. *Remedial and Special Education, 7*, 6–10.

Greenwood, C. R., Bradfield, T., Kaminski, R., Linas, M., Carta, J. J., & Nylander, D. (2011). The Response to Intervention (RTI) Approach in Early Childhood. *Focus on Exceptional Children*, *43*(9), 1–24.

Landerl K & Wimmer H. (2008) Development of word reading fluency and spelling in a consistent orthography: An 8-year follow-up. *Journal of Educational Psychology. 100(1)*:150–161.

Mastropieri, M. A., Scruggs, T. E., & Graetz, J. E. (2003). Reading comprehension instruction for secondary students: Challenges for struggling students and teachers. *Learning Disability Quarterly, 26*(4), 103-116.

Matthews, E. (2015). *Analysis of an Early Intervention Reading Program for First Grade Students.* Retrieved from http://scholarworks.waldenu.edu/cgi/viewcontent.cgi?article=1395&context=dissertations

Mislevy, R. J., Steinberg, L. S., & Almond, R. G. (2003). On the structure of educational assessment (with discussion). *Measurement: Interdisciplinary Research and Perspective,1* (1), 3-62.

Neal, R. M. (2010) ``MCMC using Hamiltonian dynamics", in the Handbook of Markov Chain Monte Carlo, S. Brooks, A. Gelman, G. L. Jones, and X.-L. Meng (editors), Chapman & Hall / CRC Press, pp. 113-162.

Nese, T. F. J., Lai, C., Anderson, D., Jamgochian, M. E., Kamata, A., Saez, L., Park, J. B., Alonzo, J., & Tinda, G. (2010). *Technical Adequacy of the easyCBM® Mathematics Measures: Grades 3-8, 2009-2010 Version* (Technical Report No: 1007). Eugene, OR: Behavioral Research and Teaching, University of Oregon.

Plummer, M. (2003). JAGS: A program for analysis of Bayesian graphical models using Gibbs sampling. *Proceeding of the 3rd International Workshop on Distributed Statistical Computing*, Viena, Austria.

R Development Core Team. (2014). R: A language and environment for statistical computing. Vienna, Austria: R Foundation for Statistical Computing. Retrieved from http://www.R-project.org

Rafferty, A. N., Brunskill, E.B., Griffiths, T. L., & Shafto, P. (2011). Faster teaching by POMDP planning. *Proceedings of the 15th International Conference on Artificial Intelligence in Education (AIED2011).* Auckland, New Zealand.

Raftery, A. E., Lewis, S. M. (1995). The number of iterations, convergence diagnostics and generic Metropolis algorithms. In: Gilks, W. R.,

Spiegelhalter, D. J., Richardson, S., eds. *Practical Markov Chain Monte Carlo.* London: Chapman and Hall.

Rock, D. A. (2007). *Growth in reading performance during the first four years in school.* (Report No: RR-07-39). Princeton, NJ: Educational Testing Service.

Ross, M. S. (1983). *Introduction to stochastic dynamic programming*. London:Academic Press.

Ross, M. S. (2000). *Introduction to Probability Models*. London: Academic Press.

Tierney, L. (1994). Markov Chain for exploring posterior distributions (with discussion). *Ann. Statist. 22*: 1701- 1762.

Tokac, Umit. (2016). Using partially observed Markov decision processes (POMDPs) to implement a response-to-intervention (RTI) framework for early reading. Doctoral Dissertation. Florida State University.

Torgesen, J.K. (2004). Avoiding the devastating downward spiral: The evidence that early intervention prevents reading failure. *American Educator, 28*, 6-19. Reprinted in the 56th Annual Commemorative Booklet of the International Dyslexia Association, November, 2005.

A Probabilistic Approach for Detection and Analysis of Cognitive Flow

Debatri Chatterjee, Aniruddha Sinha, Meghamala Sinha
TCS Research, Tata Consultancy Services Ltd.
Kolkata, India
Email: {debatri.chatterjee, aniruddha.s}@tcs.com,
meghamala.sinha@gmail.com

Sanjoy Kumar Saha
Computer Science & Engineering,
Jadavpur University,
Kolkata, India
Email: sks_ju@yahoo.co.in

Abstract

A performer may undergo a task with varying difficulty level. It is important to know the mental state in order to maintain the optimum level of performance. The mental state of an individual varies according to their IQ levels, task difficulties or other psychological or environmental reasons. We have tried to measure the cognitive state of individuals, while they are performing tasks of various complexity levels, using physiological responses like brain activation, heart rate variability and galvanic skin response. In this paper we have proposed a Bayesian network based model to probabilistically evaluate the cognitive state of an individual from the difficulty levels of the tasks, IQ level of the individual and observations made using the physiological sensing. Twenty subjects with various IQ levels are asked to play a modified Tower of London (TOL) game having three complexity levels: low, medium and high. The sensor data collected have been used to train the Bayesian model for generating the conditional probability distribution for the desired cognitive state. Results show that it can be used as a tool to determine the current cognitive state of any individual, provided we know their IQ score. In case of any contradiction between the desired cognitive state (obtained from prior knowledge) and the observed cognitive state (obtained during testing), the personal insights of a performer is analyzed.

1. INTRODUCTION

Cognitive flow is defined as a state of mind which is achieved while a person is performing a task with complete concentration and engagement. This state is closely related to the balance between the challenge of any task and the skill of the person executing the task. If the task difficulty is low compared to the held skill level of an individual, the person tends to be in the bored state (Csíkszentmihályi, Mihály, 1990). Alternatively, if the task difficulty is high, and the skill level is low, the person is supposed to be in the anxiety state. However, if the skill and task difficulty level matches, the person enter *Flow states* i.e. a state of focused concentration with a sense of enjoyment (Csíkszentmihályi, Mihály, 1997), (Csíkszentmihályi, Mihály, 1999). In our case, the skill level of an individual is directly related to his or her IQ level and is treated as the prior knowledge of the individual. On the other hand, the challenge of the task is synonymous to the task difficulty level. Hence the flow-boredom state can be represented as shown in Figure 1.

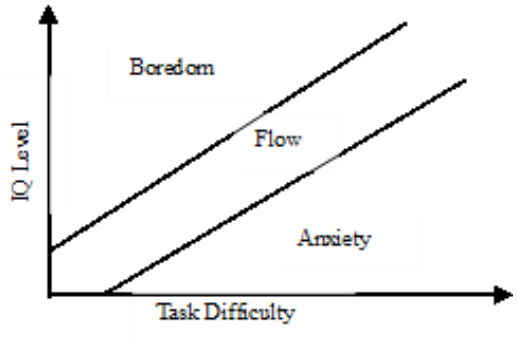


Figure 1: Flow state for IQ level and Task Difficulty

For a better learning experience, it is necessary for students to remain in the flow state throughout the session. It is a challenging task to create an environment which is enjoyable to a student and attracts complete attention as well as keep them motivated. To achieve this it is important to detect whether the student is in their flow state based on which we can induce a tailored learning environment. Present state of the art literature suggests using standard flow state questionnaires for measuring flow state. This method is indirect and might be biased. A more reliable

approach would be to estimate the flow state based on physiological changes in an individual (Sinha, Aniruddha, et al, 2015).

In this paper, our main purpose is to detect the cognitive flow with the help of physiological signals and probabilistic reasoning models. In order to do this, a modified Tower of London game (Schnirman, Geoffrey M., Marilyn C. Welsh, and Paul D. Retzlaff, 1998) is implemented for various difficulty levels in PEBL: The Psychology Experiment Building Language (Mueller, Shane T., and Brian J. Piper, 2014).

Our work mainly consists of following phases –

(i) Extraction and analysis of data from various physiological sensors like Electro-encephalogram (EEG), Heart-rate variability (HRV) and Galvanic Skin response (GSR) while a participant is performing a standard psychometric test of various difficulty levels.

(ii) Creating a Bayesian Network (BN) framework with the objective to correlate the nodes based on sensor data as mentioned in previous step with the state nodes by using data obtained in step (i).

(iii) Validation of BN based model for test data

We have used the Bayesian Network based model because it provides a theoretically efficient and consistent mechanism for processing imprecise and uncertain information. Although there has been a growing interest among researchers to use BN in the field of education and student knowledge evaluation (Chrysafiadi, K. and Virvou, M, 2013), very little work has been done to model EEG and other physiological sensor data. We have created a model for our problem domain by representing a high level probability distribution over a set of random variable denoting the different states which have direct dependencies among themselves. Conditional probability tables for each node are updated based on the sensor data.

The paper is organized as follows-. In section 2, we have given a brief summary of existing approaches adopted for analysing flow state and related works. In section 3, we have described the design of the game, experimental setup, methodologies for the analysis of sensor data and construction of Bayesian network. Section 4, presents the analysis of results that we achieved from the experiments. Finally in section 5, we have concluded and have provided the future prospect of this work.

2. RELATED WORK

The 'Flow state' has been described (Csikszentmihályi, Mihály; Harper & Row, 2015) as an experience achieved by a person while performing a task and is dependent on the personality and ability of the person. A Flow state can only be achieved when a person is engaged in an active task. Passive tasks such as watching television, taking a bath etc. do not induce flow experience (Csikszentmihályi, M., Larson, R., & Prescott, S, 1977), (DelleFave, A., & Bassi, M., 2000). The three conditions (Csikszentmihályi, M.; Abuhamdeh, S. & Nakamura, J, 2005) that must be satisfied to achieve flow are: (a) the task must have a clear goal and rate of progress, (b) the task should accompany a continuous feedback process to help the person maintain the flow state and (c) the challenge level of the task and the skill level of the person must be balanced. In the fields of education, detecting the flow state of mind is important to provide appropriate learning environment for students.

Various researches have been conducted for Flow measurement in different domains like piano playing (de Manzano, Örjan, et al, 2010), video-game (Kramer, Daniel, 2007), online games (Hsu, Chin-Lung, and Hsi-Peng Lu, 2004), social networking sites (Mauri, Maurizio, et al, 2011), e-commerce business (Koufaris, Marios, 2002) etc.

Most of these approaches use indirect methods for measuring flow (Csikszentmihalyi, Mihaly, and Isabella Selega Csikszentmihalyi, 1992), (Novak, Thomas P., and Donna L. Hoffman, 1997), (Nakamura, J. and Csíkszentmihályi, 2009). The EEG is recently being used extensively in the fields of education with the help of Brain Computer Interface (BCI) technology (van Schaik, Paul, Stewart Martin, and Michael Vallance, 2012). A greater left temporal alpha activity is noticed (Kramer, Daniel, 2007) indicating the flow state of the performer in comparison to the right temporal lobe. The mid beta and theta activity also have a distinctive effect on performance whereas no significant effect was found in the delta waveforms. Higher alpha activity coupled with lower beta activity to characterize the flow state (Mauri, Maurizio, et al, 2011). Researchers are attempting to measure boredom, anxiety etc. from EEG signals (Chanel Guillaume et. al, 2008 and Berta Riccardo et.al, 2013). Recently low cost devices are being used for analysing the effect of various elementary cognitive tasks (Chatterjee, Debatri et al., 2015). Some of these works also suggested using other physiological responses like GSR and heart rate for assessing the flow state (Chaouachi, Maher and Claude Frasson, 2010). The main problem of using multichannel physiological sensors is that, we have to find out an appropriate mechanism for fusing the results obtained from multiple sensors.

Bayesian network (Pearl, Judea, 1986) is becoming an increasingly popular technique to model uncertain and complex domains. Unlike classical statistical models, BN allow the introduction of prior knowledge into models. This prevents extraneous data to be considered which might alter desired results. BN uses the concept of conditional probability which is proven to be very useful in applications to the real world problem domain, where probability of occurrence of an event is conditionally dependent on the probability of occurrence of a previous event.

Bayesian Network modelling has been used in the areas of medicine (Koufaris, Marios, 2002), document classification, information retrieval (Luis M. de Campos,

Juan M. Fernández-Luna and Juan F. Huete, 2004), image processing, decision support system (F.J. Díez, J. Mira, E. Iturralde and S. Zubillaga, 1997*)*, gaming, bioinformatics (Neapolitan, Richard, 2009), gene analysis (Friedman, N.; Linial, M.; Nachman, I.; Pe'er, D, 2000) etc.

In the present work, BN, a desired state is derived based on the static information namely IQ of the individual and difficulty of the task. The estimated state is derived based on the sensor observation while the task is performed. The desired and the estimated state may not match as the prior characterization of individual in an exact manner is not possible. In that case, the model can help to decide the task level to maintain the flow state of the individual.

3. METHODOLOGY

This section explains the game designed and the methodology adopted for measurement and analysis of physiological signals and creation of BN model based on the findings.

3.1 GAME DESIGN

Tower of London (TOL) is a classical puzzle based game used by psychologists for assessment of executive functions and planning capabilities of an individual. We modified the standard TOL-R (Schnirmanetal, 1998) game for three levels of difficulties (Difficult, Moderate and Easy) in PEBL. For a game session, a target configuration is shown at the top of the screen as shown in fig. 2. The goal is to move a pile of disks given at the bottom so that the given assembly matches the target configuration shown on the top of the screen. Participants can only move one disk at a time, and cannot move a disk onto a pile that has no more room (indicated by the size of the grey rectangle). Participant has to click on the pile they want to move a disk off, and it will move up above the piles. Next, they click on another pile, and the disk will move down to that pile. There is a time limit to finish each game. Participants are instructed to finish each game within the allotted time. If the participant fails to finish the game within the session related time, the session ends and a new game starts. Information like start time of the game, duration, number of moves, total number of success etc. per session are stored.

The complexity of each game is determined with the help of the number of different coloured disks, the minimum number of moves needed to finish the game and the number of available empty space among the stacks for the disk to move around. Hence, the game complexity G_{com} is defined as,

$$G_{com} = \frac{N}{N_{disk} + N_{space}} \quad (1)$$

where, N is the total number of moves, N_{disk} is number of disks in the game and N_{space} and is the available number of empty space.

Screen shots for three different levels of games are shown in Figure 2. The minimum number of moves per disk for the low difficulty game is chosen as two and the total number of different disk is two, the complexity for this level of game is 3/(2+4)=0.5. Similarly, the complexities for the medium difficulty game is 4/(3+3)=0.67 and high difficulty game is 8/(4+2)=1.33 according to (1). These calculations show that the complexity for each task increases with respect to increase in difficulty level. The number of moves and number of disks for various complexity levels are chosen based on user feedback collected from 20 participants who also participated in the data collection.

The low difficulty level session consist of 10 set of games with 30 sec duration each. Similarly the medium session consist of 6 set of games each having 30 sec duration. The difficult session consist of 4 set of games with 30 sec duration each. For lower complexity game it is usually finished by most of the participants before 30 seconds, hence the number of games for various complexity levels are varied so that the completion time for easy, medium and hard sessions are comparable. Finally, it is found that for all the participants the minimum time to complete among all types (low, medium and high) of tasks is 90 seconds, hence the corresponding sensor data are considered for further processing.

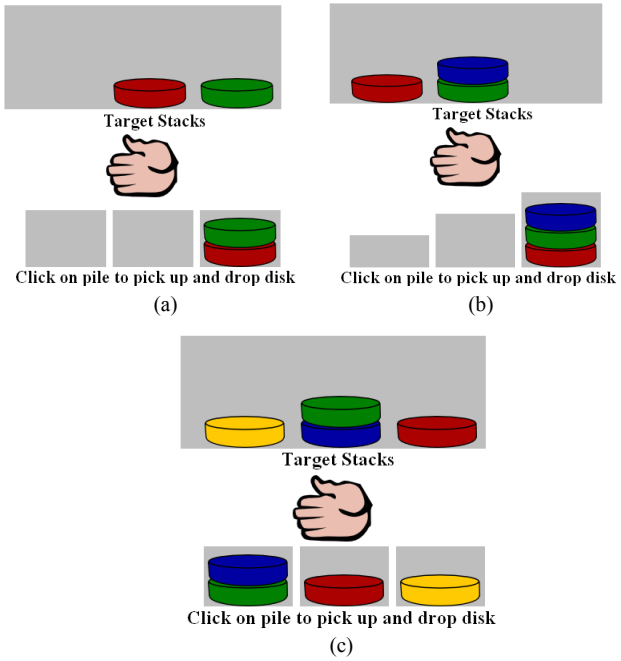


Figure 2: Screenshot of the three Tower of London games (a) Low (b) Medium and (c) High

3.2 EXPERIMENTAL SETUP

3.2.1 Participants

We have selected a group of 20 participants of varying IQ, from our research lab. The IQ scores are found to vary between 80 and 131. The average age of the participants are 22-30 years. They are all right handed male engineers belonging to similar socio-economic background. They had normal or corrected to normal vision. The selection is made to reduce the participant related bias so that the variation is only in their IQ levels.

3.2.2 Data Collection

An in house python based data capture tool is used for the data collection. The capture tool allows participants to play the game on a standard 17 inch computer screen placed at a viewing distance of approximately 25 inch and simultaneously allows to collect the EEG, GSR and Photoplethysmogram (PPG) signals. Participants are asked to play the game while wearing a single lead EEG device from Neurosky[1]. It is a dry sensor with a lead placed in FP1 position and the grounding is done with a clip fixed to left earlobe. We have used a GSR device from eSense[2] to record the variations in skin conductance level. All the participants are right handed and wore the GSR sensors on the middle and index fingers of the left hand. The right hand is kept completely free so that they can play the game comfortably using mouse. The oxygen saturation level and the pulse rate are measured by the pulse oximeter from Contec[3], through the left ring finger. The devices used are shown in Figure 3. The participants are asked to play three session of the game (High, Low and Medium) with a resting period of 5 min between each game. For half of the participants the order of the game followed is high, medium, low and for remaining half low, medium, high sequence is followed. During the game, scores, number of moves, duration etc. are also recorded for further analysis.

Figure 3: Data collection devices: (a) Neurosky EEG device (b) GSR device from eSense (c) Pulse oxymeter from Contec

3.2.3 Participant Feedback Form

The standard Game-flow indicator (GFI) feedback form has been used to assess if a participant experienced boredom and flow experiences during the experiments and the findings are used as the reference. This questionnaire based feedback form described by (Bakker, A. B, 2005) and (Bakker, A. B., 2008) is ideal to measure level of engagement while playing a game. For doing this, the overall scores for both flow and boredom questionnaires are calculated assuming 1= strongly disagree, 2= disagree, 3 = undecided, 4 = agree and 5 = strongly agree. Different questions are asked regarding the experience of participants while playing the game.

The participants are asked to fill up three feedback forms, one for each of the three games, immediately after the end of each session. This feedback is necessary as they provide a ground truth for cognitive flow along with the game data. We have used the feedback forms to analyze the contradictions which we have seen during the Bayesian Network analysis as explained in section 3.4.

3.3 SENSOR DATA ANALYSIS

3.3.1 *Skill-challenge analysis using EEG signals*

In the present work, we have experimented with various frequency band energies described by (W. Klimesch, 1999), (H. Sijuan, 2010) and time domain Hjorth parameters as described by (Gudmundsson, Runarsson, Sigurdsson, Eiriksdottir, Johnsen, 2007), (V. Carmen, et al. 2009) as shown in (2) and (3).

$$F = \{E^{\delta}, E^{\theta}, E^{\alpha}, E^{l\beta}, E^{m\beta}, H^{a}, H^{m}, H^{c}\} \quad (2)$$

The first five features are the energies in various frequency bands namely, delta (E^{δ} as 0.5 - 4 Hz), theta (E^{θ} as 4 – 7.5 Hz), alpha (E^{α} as 7.5 – 12.5 Hz), low-beta ($E^{l\beta}$ as 12.5 – 16 Hz) and mid-beta ($E^{m\beta}$ as 16 – 20 Hz) respectively. The energies in each band are extracted using Welch's power spectral density as defined by (Welch, Peter, 1967). The last three features in (2) are the Hjorth parameters namely Activity (H^{a}), Mobility (H^{m}) and Complexity (H^{c}) respectively as given in (3).

$$H^{a} = \mathrm{var}(x(t)) \quad H^{m} = \sqrt{\frac{H^{a}(\frac{dx(t)}{dt})}{H^{a}(x(t))}} \quad H^{c} = \frac{H^{m}(\frac{dx(t)}{dt})}{H^{m}(x(t))} \quad (3)$$

Here *x(t)* indicates the time domain signal in a window of duration 1 sec and $\frac{dx(t)}{dt}$ is the first order derivative of the signal.

3.3.2 *Analysis of GSR signal*

The galvanic skin response (GSR) is the electro-dermal response where the skin conductance changes with the

[1]http://neurosky.com

[2]https://www.mindfield.de/en/biofeedback/products/esense/esense-skin-response

[3]http://www.coopermedical.com/overnight-pulse-ox/cms-50d-plus-recording-fingertip-pulse-oximeter.html

amount of secretion from the sweat glands in presence of stressful, likeable events. Therefore GSR can be used as a good predictor of concentration, mental workload etc. in flow study (Nourbakhsh, Nargess, et al, 2012). During flow experience, subjects should experience a higher concentration and focus on the task. This can be measured using the GSR.

The GSR device consists of two electrodes which applies a constant voltage to the skin. The current which then flows through the skin can be detected. The GSR signal consists of two components: phasic, which is the fast varying component and tonic, which is the slow component. Both contain important information associated with specific physiological aspect of the mental states. The tonic component (T_c) is calculated only by taking the inverse transform of first few Fourier coefficients (4) and the phasic component (P_c) is calculated by inversing the higher order coefficient of the Fourier coefficients (5).The eSense GSR sensor has 5Hz sampling frequency hence we have used first 4 components (0<=k<=3) of the Fourier coefficients which corresponds to 0.5Hz. The IFFT in (4) and (5) corresponds to Inverse Fast Fourier Transform.The tonic component (T_c) of the GSR signal is computed for windows of duration 1 second. Next to investigate the distribution T_c over the task interval, we compute the mean and the kurtosis as given in (6) using N successive windows for each task (low, medium, high). These parameters provide the statistical information on the way the skin conductance changes for an individual over time, for a given task.

$$T_c = IFFT(\sum_{n=0}^{N-1} x(n).e^{-j(\frac{2\pi}{N})nk}),\ k=0,1,2,3 \quad (4)$$

$$P_c = IFFT(\sum_{n=0}^{N-1} x(n).e^{-j(\frac{2\pi}{N})nk}),\ k=4,5,...N-1 \quad (5)$$

$$T_c^{mean} = mean(T_c(n)),\ T_c^{kurt} = kurtosis(T_c(n)),\ 1 \le n \le N \quad (6)$$

3.3.3 *Calculation of HRV from PPG signals*

Stress level can be evaluated using Heart rate variability (J. Taelman, S. Vandeput, A. Spaepen and S. Van Huffel, 2009), (McDuff, Daniel, Sarah Gontarek, and Rosalind Picard, 2014). When the task challenge is low compared to the subjects' skill level, then the heart rate variability is high compared to the flow state where task challenge matches the skill. We have used the SPO2 device wearable on the ring finger for sensing the Photoplethysmogram (PPG) signal. We have calculated three time domain HRV parameters namely 1) rMSSD (root mean square of successive differences between adjacent NN intervals), 2) SDSD (Standard deviation of successive differences between adjacent NN intervals), 3) SDNN (Successive difference between NN intervals) as explained in (McDuff, 2014). Out of these SDNN is found to give a better indication of the stress level for our experiment.

3.4 BAYESIAN NETWORK CONSTRUCTION

Our problem statements for creating the Bayesian Network are as follows -

- A number of random participants are performing a task in a controlled environment.
- The task can be of three difficulty levels: easy, moderate or hard.
- The participants are categorized based on their intelligence levels i.e. the IQ scores.
- The participants can be in either of three states: flow, boredom or anxiety.
- Their performance and feedback are recorded to determine their mental state and experience.
- During data capture, three physiological sensors namely EEG sensor, HRV sensor and GSR sensor are used.

Now a Bayesian Network (BN) framework is created with the objective to diagnose and investigate the different relationships between the sensors nodes and the state nodes as shown in Figure 4.

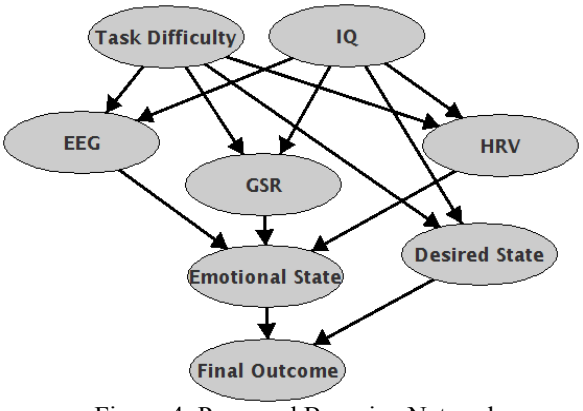


Figure 4: Proposed Bayesian Network

The BN consists of three types of nodes:

Evidence nodes: Task Difficulty (TD), Intelligence Quotient (IQ)

Sensor nodes: GSR, GSR, HRV

State nodes: Emotional State (O), Desired State (D), Final Outcome (F)

Through in the Evidence nodes we can input the static knowledge or evidence which we have in our problem domain. In our case, we know beforehand, the IQ level of the participants and also the difficulty levels of task they are given. The Evidence nodes serve as parents to the Sensor nodes, since the sensor readings are a direct causal effect of the two conditions (IQ & TD). Based on the results obtained from sensors, we can predict the current cognitive state of the participants. The derived state i.e. the Emotional state (O) from the sensor reading is compared with the Desired state (achieved from ground truth during training) and finally a conclusion (F) is drawn based on

these two states. We have used a separate desired state in order to compare it to the results derived from the signals about what emotional condition the subject is currently in.

We have created the Bayesian Network in SamIam[4], a comprehensive tool for modelling and reasoning with Bayesian networks, developed in Java by the Automated Reasoning Group of Professor Adnan Darwiche, UCLA.

At the beginning of the experiment, the participants are categorized based on their IQ levels measured by a standardized IQ[5] test prior to playing the game. This IQ test is a free online test from Brainmetrix.

The IQ level and the task difficulty level serve as evidence nodes in the BN. Depending on the state of these two nodes, the sensor readings vary. Hence all the sensor nodes have them as immediate parents. The readings of all three sensors are used to analysis the emotion state of the participants and are compared to the ground truth i.e. the desired state. If two state matches, then we can conclude the actual state via the final outcome and that the Bayesian Network is expected to give the correct outcome. If not, then there is a case of contradiction caused by one of more sensors giving faulty state. In case of a contradiction between observed state and desired state the BN model can be used to resolve this conflict by reasoning between the various nodes using probabilistic queries.

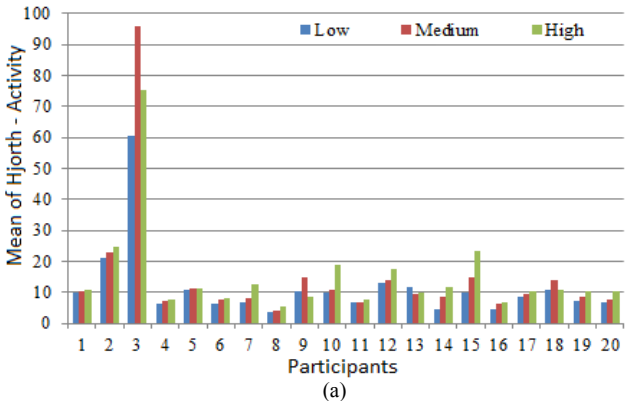
4. RESULTS AND DISCUSSIONS

4.1 EEG SIGNAL ANALYSIS

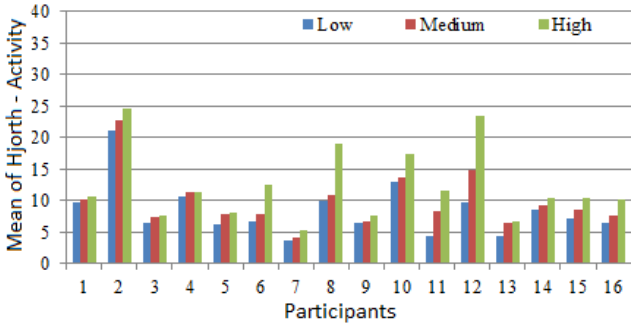
Various features of the EEG signals namely alpha, beta, theta, delta, attention, meditation, Hjorth etc. for all the experiments for all the participants are extracted using (2) and (3). Among all features the *Activity* measure of Hjorth parameter is found to be indicative of variations of brain signals with difficulty level. The raw EEG signal is used for calculating activity for windows of duration 1 sec and the overall mean is taken for each session of the game of all the participants. The results for different difficulty levels (Low, Medium and High) are combined separately for all the 20 participants and compared as shown in Figure 5(a). We have found good amount of separation between three different tasks for 16 out of 20 participants. The comparison for these selected participants are shown in Figure 5(b).

4.1 GSR SIGNAL ANALYSIS

The GSR data is subdivided into a number of windows of duration 1 sec. Next we calculate both tonic and phasic power using (4) and (5).The phasic power does not show sufficient separation between different task levels whereas the tonic gives good separation. The mean and kurtosis of the tonic power is calculated using (6). The plot of mean and kurtosis of tonic component for GSR are shown in Figure 6(a) and Figure 6(b). The line representing the medium difficulty game for all the subjects is found to overlap over the other two; the separation between the medium games across participants is not consistent. Hence for better representation, we have only plotted the GSR for high and low difficulty games.



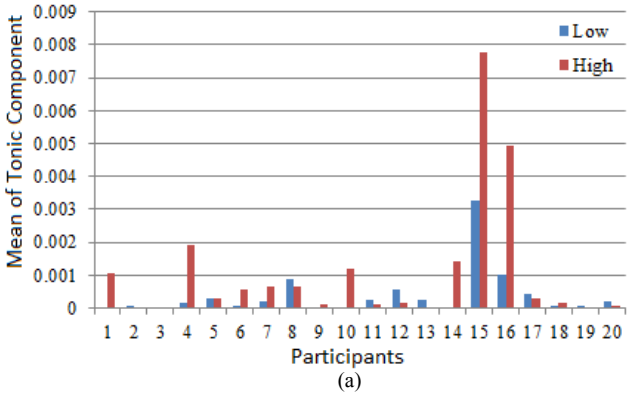
(a)



(b)

Figure 5: Separation between Low, Medium, High difficulty task for the Hjorth calculation of EEG signals for (a) 20 subjects and (b) 16 subjects

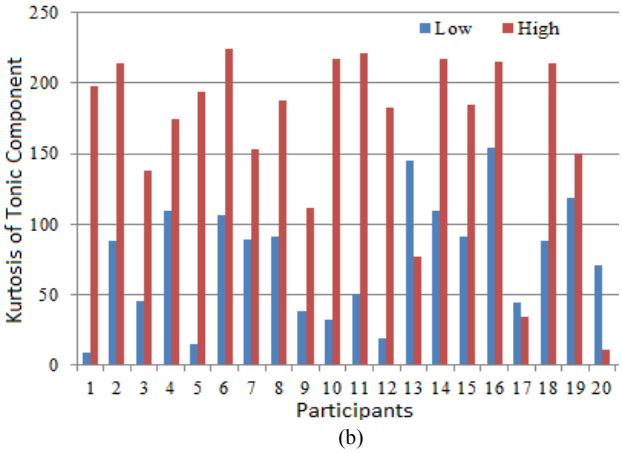
Out of 20 participants 8 participants (6, 7, 8, 9, 10, 11, 12 and 20) played the game in high-medium-low sequence and the remaining participants played in low-medium-high sequence. It is evident from the plots that kurtosis performs better than mean of the tonic power in separating the low and high difficulty tasks.



(a)

[4]http://reasoning.cs.ucla.edu/samiam/

[5]http://www.brainmetrix.com/free-iq-test/



(b)

Figure 6: Separation between the (a) Mean and (b) Kurtosis of the Tonic (μS) component of GSR during the Low and High difficulty games.

4.2 PPG SIGNAL ANALYSIS

The PPG signal analysis does not give a clear, consistent separation between difficulty levels of the task for all the participants. We have plotted the successive difference between NN intervals (SDNN) in Figure 7. For 10 out of 20 subjects there is a separation between the two SDNN values out of which only 5 have the value for the low game lower than the high game. Hence any substantial information cannot be derived from this result of PPG analysis.

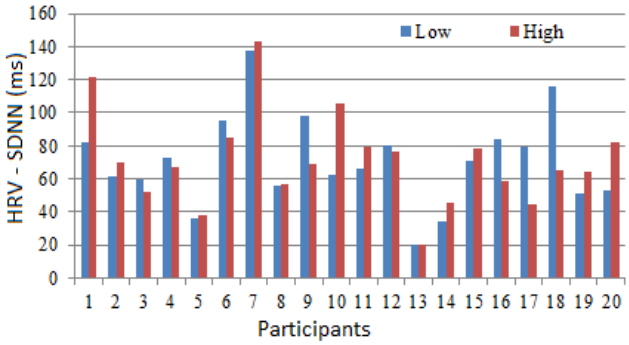


Figure 7: Separation of the HRV(SDNN) in msec for all the 20 subjects between the high and low difficulty game

4.3 BAYESIAN NETWORK ANALYSIS

The Bayesian network framework, given in Figure 4, has been trained based on the results obtained from physiological sensor data collected during the experiments. The conditional probability tables associated with each node have been updated according to the occurrence of their respective parent node.

Given a participant of a particular IQ level and solving a game of a particular difficulty level, we have collected each of the sensor data and have calculated their probability of occurrence over all possible conditions. The scores of IQ less than 89 are treated as Low IQ, scores between 90 and 109 are treated as Medium IQ and scores with 110 is treated as High IQ. In the present study we got 6 participants having high IQ, 9 participants having medium IQ and 5 participants having low IQ.

We have used the data for 18 out of 20 participants to train the BN. The sensor data for all the training participants (18 subjects) are classified into three levels (Low, Medium, High) based on their observations. An example is shown in Table 1 where the levels of the EEG sensor data are shown for 5 training participants with High IQ, playing three levels of games. It can be seen that for the easy game (TD = low), all of them have low EEG feature values, for medium game (TD = medium), one of them has high EEG feature and similarly for difficult game (TD = high). The probability of occurrence of each level is calculated and updated in the respective Conditional Probability Table (CPT) of the BN. After the training the final CPT for the EEG node is shown in Table 2. The same rule is followed for the rest of the sensor nodes as well.

Table 1: The levels for EEG Sensor data (Mean of Hjorth - Activity) for 5 Participants with High IQ and playing three levels of game (TD). Here L-Low, M-Medium, H-High.

TD	**EEG level (Mean of Hjorth - Activity)**				
L	L	L	L	L	L
M	M	H	M	M	M
H	H	M	H	H	H

Table 2: Conditional probabilities of the EEG node is shown for the pair (TD, IQ)

EEG	**(L,H)**	**(M,H)**	**(H,H)**	**(L,M)**	**(M,M)**	**(H,M)**
L	0.98	0.02	0.02	0.98	0.01	0.01
M	0.01	0.79	0.19	0.01	0.98	0.01
H	0.01	0.19	0.79	0.01	0.01	0.98
EEG	**(L,L)**	**(M,L)**	**(H,L)**			
L	0.5	0.25	0.25			
M	0.25	0.5	0.25			
H	0.25	0.25	0.5			

The data for the remaining 2 participants (having high IQ and low IQ) are used to validate whether the BN nodes are providing the correct results. Different combinations of evidences for these two subjects are checked in the BN (Figure 8 through Figure 11) and the reasons for contradictions are explained. The states of the BN are shown as rectangular blocks containing the percentage equivalent of the probability values of the random variables.

Case 1: A participant of high IQ is playing a high difficulty level game

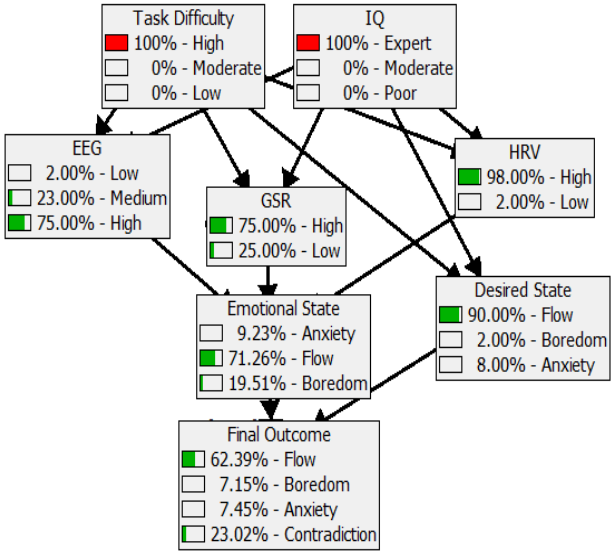


Figure 8: Bayesian Network in Query mode for the evidence Task Difficulty= High and IQ=Expert/High

In this case we can see that the Emotional State (O) matches with the Desired State (D) hence the final outcome is Flow state. Also all the sensors are providing the expected result, except the EEG feature value. The subject used in this experiment has high IQ. From the questionnaire based survey we have found that the participants has reported to be sometimes in anxious state during the high difficulty game. This could be the reason for the EEG feature value to be very high for him. We have found from participant's feedback that he was mostly in the flow state. Hence the output of the Bayesian model seems to be correlating with the participant's feedback.

Case 2: A participant of high IQ is playing a low difficulty level task

In this case we can see that if a participant with high IQ is playing a low difficulty task, the sensor nodes are providing the desired states. This is shown in Figure 9. The EEG feature value and GSR states are in low state as expected, as the participant is expected to be in a relaxed state. However, the HRV indicates a low state as opposed to the expected high state leading to the contradiction. If a highly intelligent participant is playing a very easy task with low concentration (GSR=Low) then he / she might be in a restless state which can lead to anxiety. Hence the true mental state cannot be correctly determined.

Case 3: A participant of low IQ is playing a high difficulty level task

In this case we can see that all the states are providing the expected result except the EEG feature value as shown in Figure 10. The participant in this experiment has low IQ. According to participant's feedback, he was in the flow state without being too anxious and matches with the Emotional State of the BN. However a contradiction is shown in the Final Outcome indicating a deviation of the behaviour of the participant from the expected one.

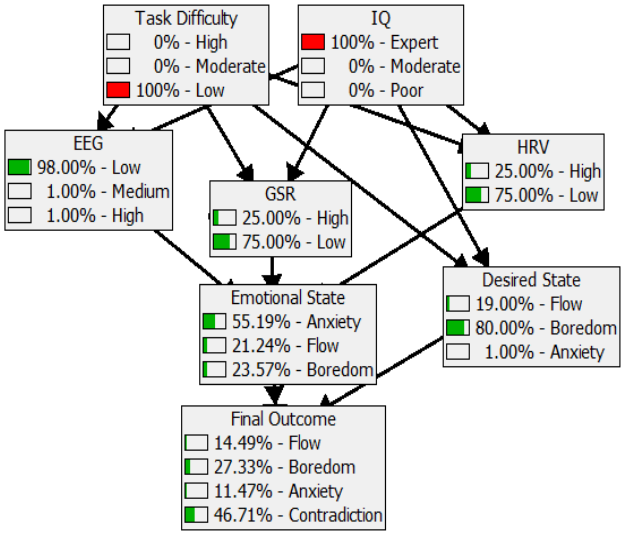


Figure 9: Bayesian Network in Query mode for the evidence Task Difficulty=Low and IQ=Expert/High

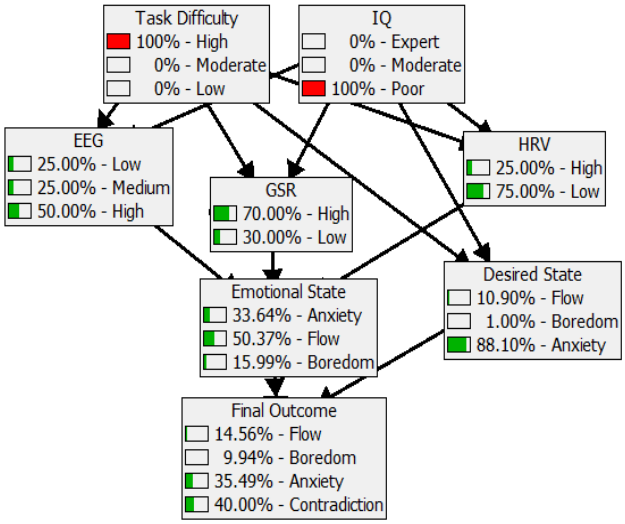


Figure 10: Bayesian Network in Query mode for the evidence Task Difficulty= High and IQ=Poor/Low

Case 4: A participant of low IQ is playing a low difficulty level task

In this case we can see that all the states are providing the expected result except the EEG feature value as shown in Figure 11. The participants used in this experiment has low IQ. We have found from participants' feedback that he was in the relaxed state and did not find the game too difficult. This could be the reason for the EEG feature value to be low but with very low percentage. It can also be observed from all the four cases that if the contradiction option is ignored in the final outcome state, then the next highest probability belongs to the state same as the desired state.

This shows that while the Bayesian Network provides the correct prediction of the actual mental state, given the sensor data and static knowledge, it also hints at the possibility of a faulty sensor data based on the probability of the contradiction.

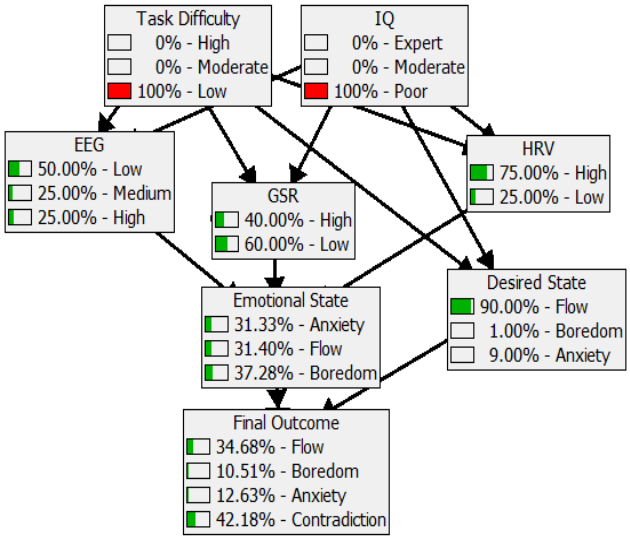


Figure 11: Bayesian Network in Query mode for the evidence Task Difficulty= Low and IQ=Poor/Low

5. CONCLUSION AND FUTURE WORK

In this work we have measured the cognitive state of a participant performing task of varying difficulty level from brain signals and certain physiological parameters like heart rate variability and galvanic skin response. Results show that EEG (Hjorth-Activity) and GSR signals can be successfully used to distinguish the performance for all the participants in three different level tasks. On the other hand, the analysis of the HRV data is not providing consistent information. These results can be fused together and analyzed along with the performance and feedback data to get further insights of the mental state for the participants.

From our proposed Bayesian Network we need to determine whether there is a contradiction between observed and desired state. In case of contradiction we need to find out whether there are any sensor(s) nodes giving faulty data. In future a large number of participants are to be considered for further analysis. Moreover, the workflow of the Bayesian Network should be dynamic for handling a large number of data. Several knowledge under uncertainty can be determined in future by using such probabilistic graphical models. In the area of education, models can be designed for generating a sequence of questions which are intermixed in difficulty level to analyze the variation in cognitive states of an individual during the assessment process.

Acknowledgements

We are thankful to the participants who have provided their time and inputs for the data collection for the experiments.

References

Bakker, A. B. (2005). “Flow among music teachers and their students: The crossover of peak experiences”. Journal of Vocational Behavior, 66, 26–44. Retrieved from http://dspace.library.uu.nl/handle/1874/10711.

Bakker, A. B. (2008). “The work-related flow inventory: Construction and initial validation of the WOLF”. Journal of Vocational Behavior, 72(3), 400–414.

Berta, Riccardo, et al. (2013). "Electroencephalogram and physiological signal analysis for assessing flow in games." Computational Intelligence and AI in Games, IEEE Transactions on 5.2: 164-175.

V. Carmen, et al. (2009). “Time Domain Parameters as a feature for EEG-based Brain-Computer Interfaces” in Neural Networks 22: 1313-1319

Chanel, Guillaume, et al. (2008). "Boredom, engagement and anxiety as indicators for adaptation to difficulty in games." Proceedings of the 12th international conference on Entertainment and media in the ubiquitous era. ACM.

Chaouachi, Maher, and Claude Frasson. (2010). "Exploring the relationship between learner EEG mental engagement and affect." Intelligent tutoring systems. Springer Berlin Heidelberg.

Chatterjee, Debatri, et al. (2015). "Analyzing elementary cognitive tasks with Bloom's taxonomy using low cost commercial EEG device." Tenth International Conference on Intelligent Sensors, Sensor Networks and Information Processing (ISSNIP). IEEE.

Chrysafiadi, K. and Virvou, M. (2013). “Student modelling approaches: A literature review for the last decade.” Expert Systems with Applications, 40(11): 4715-4722.

Csíkszentmihályi, Mihály, (1990). “Flow: The Psychology of Optimal Experience”, New York: Harper and Row, ISBN 0-06-092043-2.

Csikszentmihalyi, Mihaly. (1997). "Finding flow.".

Csikszentmihalyi, M. (1999). “If we are so rich, why aren't we happy?” American Psychologist. Vol 54(10), 821-827.

Csikszentmihályi, Mihály; Harper & Row. (2015). "FLOW: The Psychology of Optimal Experience" Retrieved 2 April 2015.

Csikszentmihalyi, Mihaly, and Isabella Selega Csikszentmihalyi, eds. (1992). “Optimal experience: Psychological studies of flow in consciousness.” Cambridge University Press.

Csikszentmihályi, M., Larson, R., & Prescott, S. (1977). “The ecology of adolescent activity and experience.” Journal of Youth and Adolescence, 6, 281-294.

Csikszentmihályi, M.; Abuhamdeh, S. & Nakamura, J. (2005), "Flow", in Elliot, A., Handbook of Competence and Motivation, New York: The Guilford Press, pp. 598–698.

deManzano, Örjan, et al. (2010). "The psychophysiology of flow during piano playing." Emotion 10.3: 301.

DelleFave, A., & Bassi, M. (2000). "The quality of experience in adolescents' daily lives: Developmental perspectives." Genetic, Social, and General Psychology Monographs, 126, 347-367.

F.J. Díez, J. Mira, E. Iturralde and S. Zubillaga (1997). "DIAVAL, a Bayesian expert system for echocardiography". Artificial Intelligence in Medicine (Elsevier) 10 (1): 59–73.

Friedman, N.; Linial, M.; Nachman, I.; Pe'er, D. (2000). "Using Bayesian Networks to Analyze Expression Data". Journal of Computational Biology 7 (3–4): 601–620.

S. Gudmundsson, P. Runarsson, T. Sigurdsson, S. Eiriksdottir, K. Johnsen, (2007). "Reliability of quantitative EEG features", in Clinical Neurophysiology 118: 21622171, August 2007.

Hsu, Chin-Lung, and Hsi-Peng Lu. (2004). "Why do people play on-line games? An extended TAM with social influences and flow experience." Information & Management 41.7: 853-868.

W. Klimesch, (1999). "EEG alpha and Theta oscillations reflect cognitive and memory performance: a review and analysis" in Brain Research Reviews 29.

Koufaris, Marios. (2002). "Applying the technology acceptance model and flow theory to online consumer behavior." Information systems research 13.2: 205-223

Kramer, Daniel. (2007). "Predictions of performance by EEG and skin conductance." Indiana Undergraduate Journal of Cognitive Science 2: 3-13.

Luis M. de Campos, Juan M. Fernández-Luna and Juan F. Huete (2004). "Bayesian networks and information retrieval: an introduction to the special issue". Information Processing & Management (Elsevier) 40 (5): 727–733.

Mauri, Maurizio, et al. (2011). "Why is Facebook so successful? Psychophysiological measures describe a core flow state while using Facebook." Cyberpsychology, Behavior, and Social Networking 14.12: 723-731

McDuff, Daniel, Sarah Gontarek, and Rosalind Picard. (2014). "Remote measurement of cognitive stress via heart rate variability." Engineering in Medicine and Biology Society (EMBC), 36th Annual International Conference of the IEEE.

Mueller, Shane T., and Brian J. Piper. (2014). "The psychology experiment building language (pebl) and pebl test battery." Journal of neuroscience methods 222: 250-259.

Nakamura, J. and Csíkszentmihályi, M. (2009). "Flow Theory and Research." In Snyder, C. R. and Lopez, S. J. eds. Oxford handbook of Positive Psychology.Oxford University Press, Oxford, 195-206.

Neapolitan, Richard (2009). Probabilistic Methods for Bioinformatics. Burlington, MA: Morgan Kaufmann. p. 406. ISBN 9780123704764.

Nourbakhsh, Nargess, et al. (2012). "Using galvanic skin response for cognitive load measurement in arithmetic and reading tasks." Proceedings of the 24th Australian Computer-Human Interaction Conference. ACM.

Novak, Thomas P., and Donna L. Hoffman. (1997). "Measuring the flow experience among web users." Interval Research Corporation 31.

Pearl, Judea. (1986). "Fusion, propagation, and structuring in belief networks." Artificial intelligence 29.3: 241-288.

Pearl, Judea, and Stuart Russell. (1998). "Bayesian networks." Computer Science Department, University of California.

Schnirman, Geoffrey M., Marilyn C. Welsh, and Paul D. Retzlaff. (1998). "Development of the Towerof London-revised." Assessment 5.4: 355-360.

Schnirman, G.M., Welsh, M.C., Retzlaff, P.D. (1998). "Development of the Tower of London-Revised." Assessment 5(4), 355 –360.

H. Sijuan, (2010). "Feature extraction and classification of EEG for imagery movement based on mu/beta rhythms", 3rd International Conference on Biomedical Engineering and Informatics (BMEI).

Sinha, Aniruddha, et al. (2015) "Dynamic assessment of learners' mental state for an improved learning experience." Frontiers in Education Conference (FIE), 2015. 32614. IEEE.

J. Taelman, S. Vandeput, A. Spaepen and S. Van Huffel, (2009). "Influence of Mental Stress on Heart Rate and Heart Rate Variability", 4th European Conference of the International Federation for Medical and Biological Engineering, vol. 22, pp. 1366-1369.

Van Schaik, Paul, Stewart Martin, and Michael Vallance. (2012). "Measuring flow experience in an immersive virtual environment for collaborative learning." Journal of Computer Assisted Learning 28.4: 350-365.

Welch, Peter. (1967). "The use of fast Fourier transform for the estimation of power spectra: a method based on time averaging over short, modified periodograms." IEEE Transactions on audio and electroacoustics: 70-73.

Improving Predictive Accuracy Using Smart-Data rather than Big-Data: A Case Study of Soccer Teams' Evolving Performance

Anthony Constantinou
Queen Mary University of London,
London, UK, E1 4NS
a.constantinou@qmul.ac.uk

Norman Fenton
Queen Mary University of London,
London, UK, E1 4NS
n.fenton@qmul.ac.uk

EXTENDED ABSTRACT

(this paper is published as extended abstract only)

In an era of big-data the general consensus is that relationships between variables of interest surface almost by themselves. Sufficient amounts of data can nowadays reveal new insights that would otherwise have remained unknown. Inferring knowledge from data, however, imposes further challenges. For example, the 2007-08 financial crisis revealed that big-data models used by investment banks and rating agencies for decision making failed to predict real-world financial risk. This is because while such big-data models are excellent at predicting past events, they may fail to predict similar future events that are influenced by new and hence, previously unseen factors.

In many real-world domains, experts comprehend vital influential processes which data alone may fail to discover. Yet, such knowledge is normally disregarded in favor of automated learning, even when the data are limited. While automation provides major benefits, these benefits sometimes come at a cost for accuracy. This study focuses on a prediction problem that has similarities to financial risk, namely predicting evolving soccer team performance. Soccer is the world's most popular sport and constitutes an important share of the gambling market. Just like in financial risk, future team performance can be suddenly and dramatically affected by rarely seen, or previously unseen, events and so both require smarter ways of data engineering and modeling, rather than just larger amounts of data.

Most of the previous extensive work on soccer has focused on results predictions based on historical data of relevant match instances. In this study we do not consider individual match results, but rather exploit external factors which may influence the strength of a team and its resulting performance. The aim is to predict a soccer team's performance for a whole season (measured by total number of league points won) before the season starts. This is an important and enormous gambling market in itself - betters start placing bets such as which team will win the title, finish in top positions, or be relegated, as soon as the previous season ends. The need for greater accuracy in such predictions has become the subject of international interest following the 2015-16 English Premier League (EPL) season when Leicester City finished top of the league, having been priced at 5,000 to 1 to do so by many bookmakers.

We use a data and knowledge engineering approach that puts greater emphasis on applying causal knowledge and real-world 'facts' to the process of model development for real-world decision making, driven by what data are really required for inference, rather than blindly seeking 'bigger' data. We refer to this as the 'smart data' approach. We use a Bayesian network (BN) as the appropriate modelling method. Based on the soccer case study, we illustrate the reasoning towards this smart-data approach to BN modeling with two subsystems:

1. A knowledge-based intervention for informing the model about real-world time-series facts; and
2. A knowledge-based intervention for data-engineering purposes to ensure data adhere to the structure of the model.

The BN model incorporates factors such as player injuries, managerial changes, team involvement in other European competitions, and financial investments relative[1] to adversaries. The BN model is based on three distinct time components:

1. Observed events from previous season that have influenced team performance;
2. Observed events during the summer break that are expected to influence team performance;
3. Expected performance for next season, accounting for the uncertainty which arises from other unknown events which may influence team performance, such as injuries.

This process is repeated for each new season, for a total of 15 seasons. This approach enabled us to provide far more accurate predictions compared to purely data-driven standard

[1] Team A may spend £20m to improve their squad, but if the average adversary spends £30m, then the strength of Team A is expected to diminish relative to the average adversary.

non-linear regression models, which still represent the standard method for prediction in critical real-world risk assessment problems, such as in medical decision analysis (Kendrick, 2014). Specifically, we demonstrate how we managed to generate accurate predictions of the evolving performance of soccer teams based on limited data that enables us to predict, before a season starts, the total league points to be accumulated. Predictive validation over a series of 15 EPL seasons demonstrates a mean error of 4.06 points (the possible range of points a team can achieve is 0 to 114). In contrast, for two different regression based methods, the mean errors are 7.27 and 7.30.

The implications of the paper are two-fold. First, with respect to the application domain, the current state-of-the-art is extended as follows:

1. This is the first study to present a model for accurate time-series forecasting in terms of how the strength of soccer teams evolves over adjacent soccer seasons, without the need to generate predictions for individual matches.

2. Previously published match-by-match prediction models (some of them include: Karlis & Ntzoufras, 2003; Rotshtein et al., 2005; Baio & Blangiardo, 2010; Hvattum & Arntzen, 2010; Constantinou & Fenton, 2012; Constantinou & Fenton, 2013b) which fail to account for the external factors influencing team strength, are prone to an error of 8.51[2] league points accumulated per team, in terms of prior belief for team strength, and for each subsequent season. Therefore, one could improve match-by-match predictions by reducing the error in terms of prior belief.

3. Studies which assess the efficiency of the soccer gambling market (Dixon & Pope, 2004; Goddard & Asimakopoulos, 2004; Graham & Stott, 2008; Constantinou & Fenton, 2013b) may find the BN model helpful in the sense that it could help in explaining previously unexplained fluctuations in published market odds.

Second, with respect to the general strategy for learning from data, we demonstrate that seeking 'bigger' data is not always the path to follow. The model presented in this paper, for instance, is based on just 300 data instances generated over a period of 15 years. With a smart-data approach, one should aim to improve the quality, as opposed to the quantity, of a dataset which also directly influences the quality of the model. We highlight the importance of developing models based on what data we really require for inference, rather than generating a model based on what data are available which represents the conventional approach to big-data solutions. With smart-data one has to have a clear understanding of the inferences of interest. Inferring knowledge from data imposes further challenges and requires skills that merge the quantitative as well as qualitative aspects of data.

For future research, we question whether automated learning of the available data is capable of inferring real-world facts such as those incorporated into the BN model presented in this paper. It may be the case that, for many real-world problems, resulting inferences will be limited in the absence of expert intervention for data engineering as well as modeling purposes. Future research will examine the capability of causal discovery algorithms in terms of realizing various real-world facts from data, and the impact various data-engineering interventions may have on the results.

Keywords: data engineering; dynamic Bayesian networks; expert systems; football predictions; smart data; soccer predictions; temporal Bayesian networks.

[2] Note that this error assumes EPL teams, and is dependent on the size of the league. For instance, the EPL consists of 20 teams and each team has to play 38 matches. Hence, the maximum possible accumulation of points is 114.

ACKNOWLEDGEMENTS

We acknowledge the financial support by the European Research Council (ERC) for funding this research project, ERC-2013-AdG339182-BAYES_KNOWLEDGE, and Agena Ltd for software support.

REFERENCES

Baio, G., & Blangiardo, M. (2010). Bayesian hierarchical model for the prediction of football results. *Journal of Applied Statistics*, 37:2, 253- 264.

Constantinou, A., Fenton, N., & Neil, M. (2012). pi-football: A Bayesian network model for forecasting Association Football match outcomes. *Knowledge-Based Systems*, 36: 322, 339.

Constantinou, A., & Fenton, N. (2013a). Profiting from an inefficient Association Football gambling market: Prediction risk and Uncertainty using Bayesian networks. *Knowledge-Based Systems*, 50: 60-86.

Constantinou, A, & Fenton, N. (2013b). Profiting from arbitrage and odds biases of the European football gambling market. *The Journal of Gambling Business and Economics*, Vol. 7, 2: 41-70.

Dixon, M., & Pope, P. (2004). The value of statistical forecasts in the UK association football betting market. *International Journal of Forecasting*, 20, 697-711.

Goddard, J., & Asimakopoulos, I. (2004). Forecasting Football Results and the Efficiency of Fixed-odds Betting. *Journal of Forecasting*, 23, 51-66

Graham, I., & Stott, H. (2008). Predicting bookmaker odds and efficiency for UK football. *Applied Economics*, 40, 99-109.

Hvattum, L. M., & Arntzen, H. (2010). Using ELO ratings for match result prediction in association football. *International Journal of Forecasting*, 26, 460-470.

Karlis, D., & Ntzoufras, I. (2003). Analysis of sports data by using bivariate Poisson models. *The Statistician*, 52: 3, 381-393.

Kendrick, M. (2014). *Doctoring Data: How to sort out medical advice from medical nonsense*. UK, Columbus Publishing.

Rotshtein, A., Posner, M., & Rakytyanska, A. (2005). Football predictions based on a fuzzy model with genetic and neural tuning. *Cybernetics and Systems Analysis*, 41: 4, 619- 630.

Scalable Joint Modeling of Longitudinal and Point Process Data for Disease Trajectory Prediction and Improving Management of Chronic Kidney Disease

Joseph Futoma
Dept. of Statistical Science
Duke University Durham, NC 27707

Mark Sendak
Institute for Health Innovation
School of Medicine
Duke University
Durham, NC 27707

C. Blake Cameron
Division of Nephrology
Duke University
Durham, NC 27707

Katherine Heller
Dept. of Statistical Science
Duke University
Durham, NC 27707

Abstract

A major goal in personalized medicine is the ability to provide individualized predictions about the future trajectory of a disease. Moreover, for many complex chronic diseases, patients simultaneously have additional comorbid conditions. Accurate determination of the risk of developing serious complications associated with a disease or its comorbidities may be more clinically useful than prediction of future disease trajectory in such cases. We propose a novel probabilistic generative model that can provide individualized predictions of future disease progression while jointly modeling the pattern of related recurrent adverse events. We fit our model using a scalable variational inference algorithm and apply our method to a large dataset of longitudinal electronic patient health records. Our model gives superior performance in terms of both prediction of future disease trajectories and of future serious events when compared to non-joint models. Our predictions are currently being utilized by our local accountable care organization during chart reviews of high risk patients.

This poster from the UAI 2016 conference was given as an invited presentation at the Bayesian Modeling Applications Workshop

Stochastic Portfolio Theory: A Machine Learning Approach

Yves-Laurent Kom Samo
Machine Learning Research Group
Oxford-Man Institute of Quantitative Finance
University of Oxford
YLKS@ROBOTS.OX.AC.UK

Alexander Vervuurt
Mathematical Institute
Oxford-Man Institute of Quantitative Finance
University of Oxford
VERVUURT@MATHS.OX.AC.UK

Abstract

In this paper we propose a novel application of Gaussian processes (GPs) to financial asset allocation. Our approach is deeply rooted in Stochastic Portfolio Theory (SPT), a stochastic analysis framework introduced by Robert Fernholz that aims at flexibly analysing the performance of certain investment strategies in stock markets relative to benchmark indices. In particular, SPT has exhibited some investment strategies based on company sizes that, under realistic assumptions, outperform benchmark indices with probability 1 over certain time horizons. Galvanised by this result, we consider the inverse problem that consists of learning (from historical data) an optimal investment strategy based on any given set of trading characteristics, and using a user-specified optimality criterion that may go beyond outperforming a benchmark index. Although this inverse problem is of the utmost interest to investment management practitioners, it can hardly be tackled using the SPT framework. We show that our machine learning approach learns investment strategies that considerably outperform existing SPT strategies in the US stock market.

This poster from the UAI 2016 conference was given as an invited presentation at the Bayesian Modeling Applications Workshop.

MDPs with Unawareness in Robotics

Nan Rong **Joseph Y. Halpern** **Ashutosh Saxena**
Computer Science Department
Cornell University
Ithaca, NY 14853
{rongnan | halpern | asaxena}@cs.cornell.edu

Abstract

We formalize decision-making problems in robotics and automated control using continuous MDPs and actions that take place over continuous time intervals. We then approximate the continuous MDP using finer and finer discretizations. Doing this results in a family of systems, each of which has an extremely large action space, although only a few actions are "interesting". We can view the decision maker as being unaware of which actions are "interesting". We model this using MDPUs, MDPs with unawareness, where the action space is much smaller. As we show, MDPUs can be used as a general framework for learning tasks in robotic problems. We prove results on the difficulty of learning a near-optimal policy in an an MDPU for a continuous task. We apply these ideas to the problem of having a humanoid robot learn on its own how to walk.

This poster from the UAI 2016 conference was given as an invited presentation at the Bayesian Modeling Applications Workshop.

Interpretable Policies for Dynamic Product Recommendations

Marek Petrik
IBM T.J. Watson Research Center
Yorktown Heights, NY 10598
mpetrik@us.ibm.com

Ronny Luss
IBM T.J. Watson Research Center
Yorktown Heights, NY 10598
rluss@us.ibm.com

Abstract

In many applications, it may be better to compute a good interpretable policy instead of a complex optimal one. For example, a recommendation engine might perform better when accounting for user profiles, but in the absence of such loyalty data, assumptions would have to be made that increase the complexity of the recommendation policy. A simple greedy recommendation could be implemented based on aggregated user data, but another simple policy can improve on this by accounting for the fact that users come from different segments of a population. In this paper, we study the problem of computing an optimal policy that is interpretable. In particular, we consider a policy to be interpretable if the decisions (e.g., recommendations) depend only on a small number of simple state attributes (e.g., the currently viewed product). This novel model is a general Markov decision problem with action constraints over states. We show that this problem is NP hard and develop a MILP formulation that gives an exact solution when policies are restricted to being deterministic. We demonstrate the effectiveness of the approach on a real-world business case for a European tour operator's recommendation engine.

This poster from the UAI 2016 conference was given as an invited presentation at the Bayesian Modeling Applications Workshop