*STAIRS 2014*
*U. Endriss and J. Leite (Eds.)*

*doi:10.3233/978-1-61499-421-3-201*

# Integration of Temporal Abstraction and Dynamic Bayesian Networks for Coronary Heart Diagnosis

Kalia Orphanou [a,1], Athena Stassopoulou [b] and Elpida Keravnou [c]

[a] *Department of Computer Science, University of Cyprus, Nicosia, Cyprus*
[b] *Department of Computer Science, University of Nicosia, Nicosia, Cyprus*
[c] *Department of Electrical and Computer Engineering and Computer Science, Cyprus University of Technology, Limassol, Cyprus*

**Abstract.** Temporal data abstraction (TA) is a set of techniques aiming to abstract time-points into higher-level interval concepts and to detect significant trends in both low-level data and abstract concepts. Dynamic Bayesian networks (DBNs) are temporal probabilistic graphical models that model temporal processes, temporal relationships between events and state changes through time. In this paper, we propose the integration of TA methods with DBNs in the context of medical decision-support systems, by presenting an extended DBN model. More specifically, we demonstrate the derivation of temporal abstractions which are used for building the network structure. We also apply machine learning algorithms to learn the parameters of the model through data. The model is applied for diagnosis of coronary heart disease using as testbed a longitudinal dataset. The classification accuracy of our model evaluated using the evaluation metrics of Precision, Recall and F1-score, shows the effectiveness of our proposed system.

**Keywords.**
temporal abstraction, temporal reasoning, Dynamic Bayesian networks, medical diagnostic models, coronary heart disease

## 1. Introduction

Temporal abstraction (TA) and Dynamic Bayesian networks (DBNs) have been gaining interest in the research community of medical-based systems. TA [1] is a knowledge-based process which creates high-level concepts from raw data interpreted over time intervals. The derived high-level abstract concepts have proved to be helpful in various clinical tasks and domains such as therapy planning, the summarization and interpretation of patient records [2].

DBNs [3] have been proposed in the literature to incorporate the explicit or implicit representation of time. They are the most widely used temporal extension of Bayesian networks, which are graphical models representing explicitly probabilistic relationships

[1] Corresponding author: Department of Computer Science University of Cyprus P.O. Box 20537 1678 Nicosia, Cyprus; E-mail address: korfan01@cs.ucy.ac.cy

among variables. DBNs are able to model stochastic processes in discrete time and they utilize a representation of a dynamic process via a set of stochastic variables in a sequence of time slices. DBNs have many applications in medicine in tasks such as medical diagnosis, forecasting, and medical decision making [4, 5]. A detailed survey on TA and DBN applied to medicine can be found in our recent work in [6].

In this paper, we present a novel approach of integrating TA techniques with DBNs. Our recent review of the relevant literature [6] indicated that both these areas have been largely used independently of each other in clinical domains. Our proposal is that they could be effectively integrated in the context of medical decision-support systems. We apply this integration in the medical domain of coronary heart disease (CHD) using as a testbed the STULONG dataset [2]. The proposed model, called 'DBN-extended' performs a CHD diagnosis on a particular patient based on the patient's medical history.

In particular, we use temporal abstraction methodologies to extract basic abstractions (i.e. state, single trend and persistence) using the finest possible granularity. The finest granularity is the smallest time interval period during which the variable state value remains the same and it can be acquired from experts' knowledge and raw data. The derived concepts are then used for DBN model development and deployment. Learning parameters and inference algorithms are applied to the constructed model.

The paper is structured as follows. In Section 2, we provide an overview of our approach and our testbed dataset. The methodology of deriving the temporal basic abstractions is described in Section 3 and the proposed DBN-extended model is introduced in Section 4. An extensive discussion of our experiments and experimental results is given in Section 5, and we conclude in Section 6.

## 2. Overview

Our goal is to integrate temporal abstraction techniques with Dynamic Bayesian networks, thus the first step is to extend the DBN network so that its nodes represent basic temporal abstractions. In order to evaluate the benefits of this integration, we developed and deployed the extended model using as a benchmark dataset, the STULONG dataset which was collected from a longitudinal study of coronary heart disease prevention. Examples of CHD events are: acute coronary syndrome, myocardial infarction, angina pectoris and ischemic heart disease. The target group includes 1428 men who may have had, or not, a CHD event before the beginning of the study.

### *2.1. System Overview*

Our approach consists of four main phases:

i) Data preprocessing and feature selection
ii) Derivation of basic temporal abstractions (state, trend and persistence TAs) from raw data
iii) Construction of the 'DBN-extended' model and
iv) Evaluation of the model

[2] The data resource is on: http://euromise.vse.cz/challenge2004/ [Date accessed: 15 May 2014]

The first main phase consists of the feature selection process and the selection of the temporal range of observations. The selected time period is the total number of years of observations based on which the temporal abstractions were generated and the total number of time slices for the DBN were selected. We base our selection of features on the domain knowledge that we acquired from a CHD expert. The selected features are either direct or indirect risk factors (RFs) of CHD. Direct RFs include age, hypertension, cigarette smoking status (current smoker or not), dyslipidemia levels (such as Total cholesterol/HDL ratio, LDL and triglycerides levels), obesity, diabetes and history features (such as past personal history and family history). Indirect RFs include medicines treating high cholesterol (taken or not), diet (if they follow any diet or not) and exercise (if they regularly exercise or not).

The key problem for model construction is the choice of the total observation period for all patients since it ranges from 1 to 24 years. In order to remove as few records as possible from the dataset, the temporal range is chosen to be 24 years. For patients whose total observation period is less than 24 years, the CHD event is considered unknown on the years beyond their observation period. The patients' health condition is assumed to remain stable during any time period that their examination results are unknown either because their total observation period is less than 24 years or they did not take any examination at the particular time period. The target group was reduced by removing records of patients with less than three years of observations since in this study we are going to focus on the temporal aspect of the data, utilizing the advantage of long-term observations. The final target group consists of 849 individuals from whom 254 had an event at some point in time during their whole monitoring examination period.

## 3. Basic Temporal Abstractions

Temporal abstraction techniques are divided into two categories: basic and complex TAs. In this study we are concerned with basic temporal abstractions techniques such as: states, trend and persistence. One of the assumptions used in deriving temporal abstractions (state and trend) is that the abstraction value of a variable with missing raw values at any time within the interval period, is defined to be the same as its last known value. The same applies for cases when no record is defined during the required time interval.

### *3.1. State TAs*

The state abstractions determine the state of an individual parameter over a particular time period based on predefined categories. The state categories for the selected features (variables) are defined by clinical experts rules. For example, poor-controlled and well-controlled hypertension are state TAs of systolic and diastolic blood pressure values. The hypertension variable is defined by the 'poor-controlled' state label if the patient has a history of hypertension and his systolic or diastolic blood pressure levels are above the standard limits; and by the 'well-controlled' state label when a patient has a history of hypertension and his systolic or diastolic blood pressure levels are normal. Otherwise, it is defined by the 'no hypertension' label. Dyslipidemia is a state TA of dyslipidemia values. It is defined as 'Present' when a patient has any of the dyslipidemia values higher than the standard limits and 'Absent' otherwise. State TAs for the Age variable are derived using three state categories labels: a) 'Normal' when the patient is under 50 years old, b) 'High' when the patient is between 50 and 60 years old and c) 'Very High' when the patient is over 60 years old. State TAs for the rest of the variables are derived in a similar manner. All state TAs are displayed in **Table 1**.

**Table 1.** State TAs variables and their state values. Variable code is the variable name in the DBN model

| Variable | Variable Code | Value=1 | Value=2 | Value=3 |
|---|---|---|---|---|
| Smoking | Smoking | No Smoker | Current Smoker | |
| Cholesterol Medicines | medCH | Taken | Not taken | |
| Hypertension | HT | No Hypertension | Well Controlled | Poor Controlled |
| Dyslipidemia | Dislipidia | Absent | Present | |
| Obesity | Obesity | Absent | Present | |
| Age | AGE | Normal | High | Very High |
| Diet | DIET | Following Diet | Not Following Diet | |
| Exercise | Exercise | Exercising | Not Exercising | |

### 3.2. Trend TAs

Trend abstractions of a feature are generated by observing the changes between their values. Examples of trend values are: decreasing, steady and increasing. In our approach, trend abstractions of a variable are generated by comparing two or more consecutive feature values (during the interval period of 3 years) as follows: taking into consideration the trends of all the feature values of all the examinations during a particular time period interval (3 years duration - 1-3 examinations), the most frequent trend value is selected for the corresponding feature for that period. We have also used a combination of trends and state abstractions in order to define the ratio of change of a particular variable based on its state value. Trend abstraction values are:

- 'Abnormal' when the variable state value is abnormal and its trend ratio is increasing or steady
- 'AbnormalDecr' when the variable state value is abnormal and its trend ratio is decreasing
- 'NormalInc' when the variable state value is normal and its trend ratio is increasing and
- 'Normal' when the variable state value is normal and its trend ratio is decreasing or steady

The resulting trend abstractions are displayed in **Table 2**.

### 3.3. Persistence TAs

Persistence TA techniques derive maximal intervals for some property by applying persistence rules both backwards and forwards in time from the specific time of the given property. Such examples are Diabetes, Family History (FH) and the past personal history of a patient for a CHD event (HistoryEvent). For example, when someone was diagnosed with diabetes at time $t$, diabetes is present from time $t$ and onwards. Similarly, when

**Table 2.** Trend TAs variables and their trend values. Variable code is the variable name in the DBN model

| Variable Name | Variable Code | Value = 1 | Value = 2 | Value = 3 | Value =4 |
|---|---|---|---|---|---|
| Total Cholesterol/HDL Ratio | TCH/HDL | Abnormal | AbnormalDec | NormalInc | Normal |
| LDL | LDL | Abnormal | AbnormalDec | NormalInc | Normal |
| Triglycerides | TRIG | Abnormal | AbnormalDec | NormalInc | Normal |
| HDL | HDL | Increasing | Steady | Decreasing | |
| Total Cholesterol | TCH | Increasing | Steady | Decreasing | |

someone was diagnosed with a CHD event at time $t$, he has a history of event from $t + 1$ and onwards, thus the value of HistoryEvent variable is ‘Present’ from $t + 1$ until the end of the monitoring process. The FH is an example of persistence TA for the whole representation time period, since its value does not change through time. The resulted persistence TAs are displayed in **Table 3**.

**Table 3.** Persistence TAs variables and their persistence values. Variable code is the variable name in the DBN model

| Variable Name | Variable Code | Value = 1 | Value = 2 |
|---|---|---|---|
| Diabetes | Diabetes | Present | Absent |
| Past Personal History | HistoryEvent | Present | Absent |
| Family History | FH | Present | Absent |

## 4. Constructing the Dynamic Bayesian Network

The construction of the extended Dynamic Bayesian network consists of two steps: i)Building the network structure (qualitative part) and ii)Learning the parameters of the network (quantitative part).

### *4.1. Network Structure*

The network structure, as displayed in **Figure 1**, was designed by incorporating prior information elicited from medical experts and medical literature. The derived basic temporal abstractions described in Section 3 form the nodes (variables) of our DBN. The DBN framework enables us to combine all the observations of a patient as evidence and derive a probability for the hypothesis that the patient is diagnosed with a CHD, given the total evidence gathered.

The model consists of 17 variables of which 15 are observed and two are hidden (with unknown value). Hidden variables are the class variable CurrentEvent representing the diagnosis of a CHD event and the Dislipidia node. Both of these variables take two values: ‘Present’ and ‘Absent’. Dislipidia is introduced as a common effect node of TCH/HDL, LDL and triglycerides, which are direct risk factors to the class variable, using the parent divorcing method in order to simplify the parameters estimation process [7]. The variable FH is not repeated since it was modeled only as an initial condition and it is not changing over time. It is therefore shown in the network of **Figure 1**, to be outside the temporal plate. The arcs in the network are carriers of the causal and temporal relationships among the variables. Intra-slice arcs represent the static relationships among variables within the same time-slice whereas inter-slice arcs connect nodes between different time slices to represent the changes over time among the variables. The single digit numbers on the arcs denote the temporal delay of the influence of the cause node to the effect. For example, an arc labeled as 1 between the variables History of CHD (HistoryEvent) and itself denotes an influence that takes one time step which is reflected to the next time slice. On the other hand, the arc without label connecting the CHD risk factors (hypertension, obesity, etc.) to the CHD event, denotes a static influence at the same time slice.

The first time slice in the network represents the time period starting from the patients' entry examination and ending three years after their entry examination. This fixed three-year granularity is chosen as it is the finer granularity, because DBNs are not able to represent irregular time periods.

### 4.2. Learning Parameters

Having defined the structure of the network, we need to define the conditional probabilities which quantify the arcs of the network. More specifically, we need to define the prior probability for the root nodes such as: AGE, Diet, Exercise, Smoking and FH as well as the conditional probability distributions for all non-root nodes. Each table gives the conditional probability of a child node to be in each of its states (values), given all possible parent state combinations. All of the parameters are learned from data using the expectation maximization (EM) algorithm [8].

Once the network structure is defined and the network is quantified with the learned conditional probability tables, the next step is to predict the probability of the class node CurrentEvent. Each variable in the network is instantiated by the corresponding feature value. The DBN is unrolled for eight time slices in order to represent the total observation period (24 years) of all patients included in the training dataset. Then it performs inference and derives the belief in the class variable, i.e. the posterior probability of the class at $t$ to take on each of its values given the evidence (features) observed at the previous time step $t-1$ and at the current time-step.

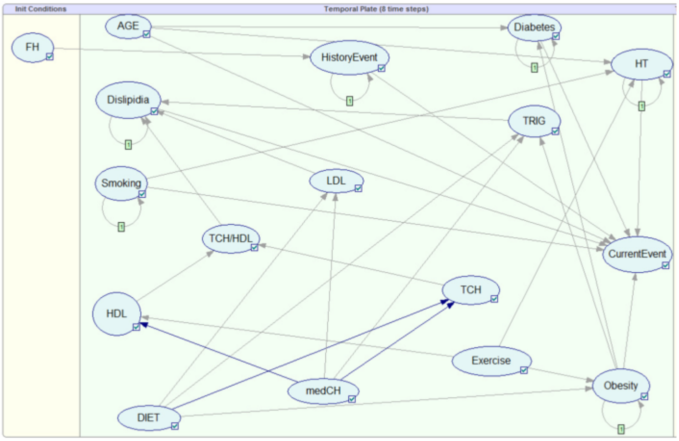


**Figure 1.** The graphical structure of the developed DBN model displaying the nodes on one time slice and temporal arcs representing the static or temporal relationships among variables.

## 5. Experimental Results and Analysis

In this section, we present the experiments performed in order to apply our methodology and evaluate the performance of our 'DBN-extended' model. For the evaluation of the accuracy of our model, we divided the dataset into training and testing using the cross-validation technique. The version of cross validation that we used in the experiments is 10 times 10-fold cross-validation, i.e. we averaged 10 runs of 10-fold cross-validation with different 10 folds in each run [9]. The classification accuracy of the model is estimated on $t = 7$ which is the last time slice.

### *5.1. Training in the Presence of Class Imbalance*

One of the most important problems in the data mining field is to deal with imbalanced datasets. The datasets present a class imbalance when there are many more examples of one class (majority class) than of the other (minority class). It is usually the case that this latter class, i.e. the unusual class, is the class of interest. Because this unusual class is rare among the general population, the class distributions are very skewed.

In the current dataset, individuals who were not diagnosed with a CHD event at a particular time period are many more than those who were diagnosed with a CHD event (minority class). Most existing classification methods tend not to perform well on minority class examples when the dataset is extremely imbalanced. One approach to tackle the problem of an imbalanced dataset is to use resampling to modify the datasets [10, 11]. This is achieved by either removing examples from the majority class (undersampling) or adding more examples to the minority class (oversampling) or a combination of both. In our system, we evaluate our classifier on two oversampling methods as well as on a combination of oversampling with undersampling. More specifically, we apply the following resampling methods: a)SMOTE-N (Synthetic Minority Oversampling Technique for nominal features) [12], which generates synthetic examples to be added to the minority class, b)random oversampling where minority cases are randomly chosen for duplication until the ratio of majority to minority reaches a desirable level and c)SMOTE-N oversampling on the minority class with random undersampling the majority class.

We performed 4 experiments, based on resampling at various ratios. **Table 4** shows the number of patients records with a CHD event ('Present') at $t = 7$ and the number of patients records without a CHD event ('Absent') in each of the 4 training data sets:

- Dataset $D1$ is the original dataset (no resampling)
- Dataset $D2$ is defined by random oversampling the minority class
- Dataset $D3$ is obtained via oversampling using SMOTE-N and finally
- Dataset $D4$ is derived using a combination of oversampling with SMOTE-N and random undersampling.

We constructed four networks, one for each experiment. The networks had the same structure but differed in their parameters, i.e. prior probabilities and the conditional probability tables. Each time a new training dataset was introduced, new network parameters were derived using training on the new set. Throughout the remaining of the paper we will refer to the four models as: $D1$, $D2$, $D3$ and $D4$. The models presented in this paper were created and tested using the SMILE inference engine and GeNIe [3].

[3] A development environment for reasoning in graphical probabilistic models, available at: http://genie.sis.pitt.edu/. [Date accessed: 15 May 2014]

### 5.2. Testing the System

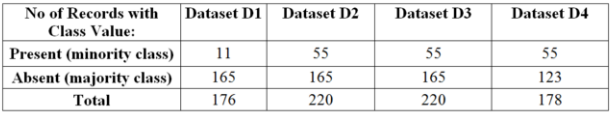
Two metrics that are commonly applied to imbalanced datasets to evaluate the performance of the models is recall (Eq 1) and precision (Eq 2). These two metrics are summarized into a third metric known as the $F_1$ measure (Eq 3). The $F_1$ measure is the combination of precision and recall which measures the effectiveness of classification in terms of a ratio of the weighted importance of recall and precision. In the evaluation of the proposed approach, both metrics are given equal importance. Recall and precision should be close to each other, otherwise the $F_1$ measure yields a value closer to the smaller of the two. Applied to our problem, precision is the ratio of the number of patients who had a CHD event at time $t$ and are correctly classified, divided by the total number of patients who are classified of having a CHD event at time $t$. Recall, on the other hand, is the ratio of the number of patients who had a CHD event at time $t$ and are correctly classified, divided by the number of patients with an actual CHD event at $t$.

$$Precision = \frac{\text{True Positive}}{\text{True Positive} + \text{False Positive}} \tag{1}$$

$$Recall = \frac{\text{True Positive}}{\text{True Positive} + \text{False Negative}} \tag{2}$$

$$F_1 score = 2 \times \frac{Precision \times Recall}{Precision + Recall} \tag{3}$$

**Table 4.** Datasets used for four experiments with and without resampling



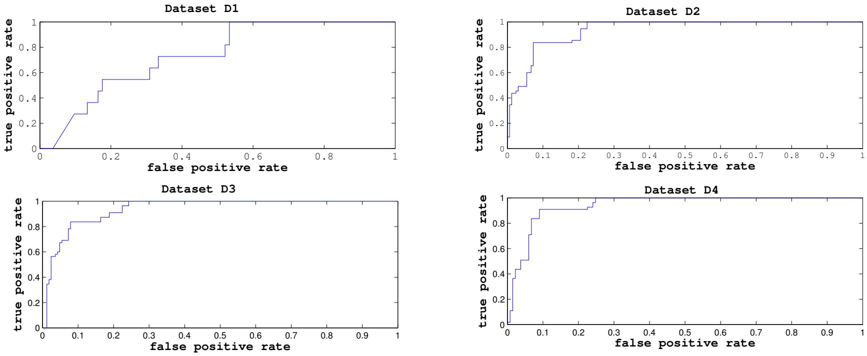
| No of Records with Class Value: | Dataset D1 | Dataset D2 | Dataset D3 | Dataset D4 |
|---|---|---|---|---|
| Present (minority class) | 11 | 55 | 55 | 55 |
| Absent (majority class) | 165 | 165 | 165 | 123 |
| Total | 176 | 220 | 220 | 178 |

The values of recall, precision and $F_1$-measure obtained from the evaluation of our model for each of the four training datasets are given in **Table 5**. As expected, the performance of the model without resampling ($D1$) is very low as this dataset is highly imbalanced and the classifier is biased towards the majority class. By applying both the random oversampling and SMOTE-N methods, we obtained dramatically improved results compared to $D1$. One risk with random oversampling, is overfitting due to placing exact duplicates of minority examples from the original set and thus making the classifier biased by remembering examples that were seen many times. The SMOTE-N technique overcomes this risk by creating synthetic examples by interpolating pairs of the closest neighbors in the minority class and introduces some new cases not included in the original dataset. The dataset derived by applying SMOTE-N oversampling combined with undersampling ($D4$) had the best classification performance. With this dataset, recall reaches as high as 91% whereas precision reaches as high as 75% yielding a combined F1-score of 82%.

**Table 5.** The evaluation results for all the four training datasets

| Evaluation Metrics | Dataset D1 | Dataset D2 | Dataset D3 | Dataset D4 |
|---|---|---|---|---|
| **Precision** | 0.176470588 | 0.647058824 | 0.680555556 | 0.746268657 |
| **Recall** | 0.272727273 | 0.8 | 0.859649123 | 0.909090909 |
| **F-score** | 0.214285714 | 0.715447154 | 0.759689922 | 0.819672131 |

We also used Receiver Operating Characteristic (ROC) curves [13, 14] to show graphically the classification performance of the four models. ROC displays graphically the trade-off between true positive rate (TPR) and false positive rate (FPR) of a classifier. TPR is the fraction of positive examples predicted correctly whereas FPR is the fraction of negative examples predicted as positive. A point on the ROC curve represents the FPR and TPR associated with the classification based on a given discrimination threshold. The threshold refers to the cut-off value above which a record is classified as positive. By varying the threshold we produce different points on the ROC curve (i.e. different (FPR,TPR) pairs). A good classification model is located as close as possible to the upper left corner of the diagram, i.e. point (TPR =1, FPR=0). The resulted graphs are displayed in **Figure 2**.


Dataset D1

Dataset D2

Dataset D3

Dataset D4

**Figure 2.** ROC curves $t = 7$ for all the four datasets.

## 6. Conclusions and Future Work

In this paper, we represented a new approach of integrating temporal abstraction with Dynamic Bayesian networks in the context of coronary heart disease. The benefits of applying our developed DBN-extended model to the CHD domain are that this extension can handle incomplete evidence in predicting disease outcomes and dealing with uncertainty which are the most usual challenges in the domain of CHD.

During our training and evaluation stages we addressed the class imbalance problem on the dataset. We have used three techniques of resampling, random oversampling, SMOTE-N and combination of SMOTE-N with undersampling to deal with the imbalance problem and developed four models by training on four different datasets. The high classification accuracy results proves the effectiveness of our proposed methodology. Our recall and precision were reaching as high as 91% and 75% respectively by applying SMOTE-N technique in combination with undersampling to the original dataset. The classification results provide a promising direction for future work. The next step is to apply the proposed approach for prognosis in the CHD domain. Estimating CHD risk for future time periods (prognosis) can help clinicians provide treatment decisions to patients that will prevent CHD events.

In addition, we are currently investigating the introduction of complex temporal abstractions to the nodes of the DBN-extended model. Complex temporal abstractions define a combination of basic TAs and/or temporal patterns. Through their representation into the DBN-extended model, we will also introduce the representation of new temporal dependencies between the variables (such as ‘meets’, ‘overlaps’ and ‘starts’)and the representation of events occuring at irregular time periods into the time slices.

## References

[1] Y. Shahar, M. A. Musen, Knowledge-based temporal abstraction in clinical domains, Artificial intelligence in medicine 8 (3) (1996) 267–298.

[2] N. Lavrač, I. Kononenko, E. Keravnou, M. Kukar, B. Zupan, Intelligent data analysis for medical diagnosis using machine learning and temporal abstraction, AI Communications 11 (3,4) (1998) 191–218.

[3] K. P. Murphy, Dynamic bayesian networks: representation, inference and learning, Ph.D. thesis, University of California (2002).

[4] Y. Xiang, K.-L. Poh, Time-critical dynamic decision making, in: Proceedings of the Fifteenth Conference Annual Conference on Uncertainty in Artificial Intelligence (UAI-99), Morgan Kaufmann, San Francisco, CA, 1999, pp. 688–695.

[5] T. Charitos, L. C. van der Gaag, S. Visscher, K. A. M. Schurink, P. J. F. Lucas, A dynamic bayesian network for diagnosing ventilator-associated pneumonia in ICU patients, Expert Systems Applications 36 (2) (2009) 1249–1258.

[6] K. Orphanou, A. Stassopoulou, E. Keravnou, Temporal abstraction and temporal bayesian networks in clinical domains: A survey, Artificial Intelligence in Medicine 60 (3) (2014) 133 – 149.

[7] F. V. Jensen, An introduction to Bayesian networks, Vol. 210, UCL press London, 1996.

[8] T. Moon, The expectation-maximization algorithm, Signal Processing Magazine, IEEE 13 (6) (1996) 47–60.

[9] J. M. Pena, J. Björkegren, J. Tegnér, Learning dynamic bayesian network models via cross-validation, Pattern Recognition Letters 26 (14) (2005) 2295 – 2308.

[10] A. Estabrooks, T. Jo, N. Japkowicz, A multiple resampling method for learning from imbalanced data sets, Computational Intelligence 20 (1) (2004) 18–36.

[11] A. Stassopoulou, M. D. Dikaiakos, Crawler detection: A bayesian approach, in: Proceedings of the International Conference on Internet Surveillance and Protection, ICISP ’06, IEEE Computer Society, Washington, DC, USA, 2006, pp. 16–.

[12] N. V. Chawla, K. W. Bowyer, L. O. Hall, W. P. Kegelmeyer, SMOTE: Synthetic Minority Over-sampling Technique, Journal of Artificial Intelligence Research 16 (2002) 321–357.

[13] T. Fawcett, An introduction to ROC analysis, Pattern Recogition Letters 27 (8) (2006) 861–874.

[14] P.-N. Tan, M. Steinbach, V. Kumar, Introduction to Data Mining, (First Edition), Addison-Wesley Longman Publishing Co., Inc., Boston, MA, USA, 2005.