Journal of Computer Science

Original Research Paper

Fuzzy Method for Online Learning of Bayesian Network Parameters

Mariana D.C. Lima and Silvia M. Nassar

Department of Informatics and Statistics, Universidade Federal de Santa Catarina, Florianopolis/SC, Brazil

Article history
Received: 01-08-2018
Revised: 08-10-2018
Accepted: 13-03-2019

Corresponding Author:
Mariana D.C. Lima
Department of Informatics and
Statistics, Universidade Federal
de Santa Catarina,
Florianopolis/SC, Brazil
Email: selvarajr@biust.ac.bw

Abstract: In learning problems, there are situations where training data is not fully available at the learning time. They are incrementally generated by time, defining a type of domain called online that has among its characteristics the possibility of data failure or even missing data. In Bayesian networks, learning is divided into two categories: structure (related to the graph of conditional relations) and parameters (related to the strength of conditional relations). In this work we present an online parameter learning method that quickly adapts to changes in the environment aiming not only the reproduction of the probability distribution (generative learning) but also the increase of accuracy in the network (discriminatory learning). Our approach is compared with the Adaptative Voting EM method considering two simulation conditions: when distributions are unknown and when distributions undergo abrupt changes. The proposed method achieves good results in both situations by adjusting to environment changes more quickly and by simplifying the parameterization of the traditional approach.

Keywords: Parameter Learning, Bayesian Networks, Online Learning, Discriminative Learning, Generative Learning

Introduction

Bayesian Networks (BN) have become extremely popular in the last decades because they have been able to map the between variables Friedman *et al*. (1997). In addition, they are an appropriate language with efficient resources for representing the joint probability distribution over a set of random variables. The technique is even more attractive by being able to model real world problems and by the interpretation of the network by non-specialists Zhou (2015).

The learning process in Bayesian Networks is divided between structure learning and parameters learning (Kurihara *et al*. (2001); Chen *et al*. (2001); Zhang and Liu (2008). While the first aims to build the network graph, the second focuses on updating the conditional probabilities among the variables.

Parameters learning algorithms are divided into two main categories: generative and discriminative Su *et al.* (2008). The first one creates conditional probabilities considering the data distribution and the second does it with the objective of increasing the accuracy on the network. Among the most used generative algorithms is the maximization of likelihood (MLE) obtained directly from the dataset and the (EM) Expectation-Maximization Dempster *et al*. (1977) algorithm in case of missing data.

One of the difficulties in parameters learning is the computational complexity of the algorithms, since the problem in the worst case is NP-hard Ratnapinda and Druzdzel (2015). There is also the risk of the algorithm being stopped at a local maximum Myers *et al*. (1999).

Online parameter learning is usually accomplished through adaptations in the generative methods by informing the influence of future data against past data. The goal of those methods is the model convergence, that is, to reproduce the distribution of the data in the Conditional Probability Tables (CPT).

Although Bayesian reasoning is probabilistic, it is possible to combine complementary techniques and reasoning in BN. Take for example, the Fuzzy-Bayesian model Brignoli (2013) that combines diffuse (fuzzy) reasoning with the probabilistic reasoning.

In a previous work e Lima (2014), a discretization method was developed for Bayesian networks through rules of data-based cuts and the overall optimization of them using genetic algorithm.

By combining different reasoning techniques on the uncertainty it is possible to address more than one face

Science Publications

of the same problem. In this work it is proposed a method that performs the online parameters learning in a hybrid way between the discriminative approach and the generative approach. The proposed method is based on the Voting EM Cohen *et al.* (2001b) and it inherits some of its characteristics as online learning and the possibility of missing data during learning.

Although there are other methods of parameters learning in literature that make the hybridism between the discriminative approach and the generative one, it is usually done by separating the variables into two distinct sets. The first set is treated in a generative way (where the goal is to reproduce the data distribution) and the second in a discriminatory manner (where the goal is increase the accuracy in classification problems). The proposed method carries out this hybridism in an integrated way between two approaches: The same variable is learned simultaneously in a generative and discriminatory way through a fuzzy learning system.

Related Work

Models in Bayesian networks are made from the graph topology and the conditional probabilities between the variables. The definition of these two properties is accomplished through the process known as learning where:

- Structure Learning: It determines whether or not there is independence between the variables of BN and it gives a score for each candidate structure
- Parameter learning: It is related to the estimation of conditional probabilities among the variables

Parameter learning is related to the estimation of Conditional Probability Tables (CPT) and it is divided into two approaches: Generative and discriminative.

In generative learning the conditional probabilities are computed directly from data. On the other hand, in the discriminative approach, learning is done considering the conditional probability of the variables in order to provide the increase of accuracy in the BN Su *et al.* (2008).

Generative learning is made from the distribution of data and it seeks the likelihood maximization Zhou (2015). The most common method is the (MLE) Maximum Likelihood Estimation for cases where there is no missing data Friedman *et al.* (1997).

When there is occurrence of missing data, the EM method is the most used Dempster *et al.* (1977). It enables the estimation of parameters through a repeating structure that toggles between two steps: E-step and M-step until it reaches the convergence. Reaches convergence some variations of EM method were proposed in the literature, for example the EM (η) Bauer *et al.* (1997). This algorithm defines the concept of learning rate in EM and the update rules considering a Bayesian network. The Voting EM method is an online version of EM (η) Cohen *et al.* (2001b), Cohen *et al.* (2001a). The main features of Voting EM are:

- Adaptation to data distribution changes
- Ability to escape from the local maximum in the likelihood function
- It reaches the convergence more rapidly than the MLE method
- Faster adaptation in cases where there are changes in data distribution when it is compared to MLE

Another method related to EM is the EM-like proposed by Saloj¨arvi *et al.* (2005). The method is a discriminative version of EM and it aims to maximize conditional probabilities rather than likelihood probabilities as it happens in classical EM method.

The pioneering method in discriminative approach is the (ELR) Extension to Logistic Regression proposed by Greiner and Zhou (2002), where CPT are estimated by a process that uses the downward gradient as a way to maximize the conditional probability. The authors show that discriminative learning requires fewer training instances than generative to converge and that usually leads to a more efficient classifier. However, the computational cost can be significantly higher.

Raina *et al.* (2003) propose a hybrid method between the generative and discriminative approach. The method divides the variables into two groups: Discriminative and generative. Therefore, if a variable has a direct influence on classification, it is learned in a discriminatory way and, if not, in a generative way. The method obtained a high accuracy rate and a low error when compared when it I compared to ELR.

Kang and Tian (2006) propose the HBayes-NB which is a hybrid approach to learning parameters and structure. The HBayes-NB performs the relaxation of the na¨ıve Bayes topology by creating additional arcs in the graph. The variables are separated into two sets: discriminative and generative. Discriminative learning is done by the ELR method and the generative by the MLE. The method obtained good results when it is tested on public databases and compared with state-of-the-art methods in classification problems.

Liu and Liao (2008) propose an online learning method made by combining MLE and VotingEM. The method proposed by the authors changes the VotingEM learning rate proportionately to the time of arrival of the data in a similar way to the MLE method. The method proposed obtained similar results to VotingEM but proved less sensitive to the parameters configuration

Su *et al.* (2008) propose the (DFE) Discriminative Frequence Estimate that learns parameters in a discriminatory way considering the data frequency. DFE is a variation of the MLE method and uses the error (loss) as a penalty in learning. The method was compared with MLE, ELR and with an ensemble method in several public databases of the UCI repository. The DFE obtained good results and the authors conclude that the method is computationally efficient, converges quickly and has results similar to the state-of-the-art methods.

Pernkopf and Wohlmayr (2009) propose three discriminative methods of parameters learning. The first is an extension to RB of the Baum-Welch algorithm Bridle (1990). The other two methods are based on EM-like Saloj¨arvi *et al*. (2005): (ECL) Exact conditional likelihood method and (ACL) Approximate Conditional likelihood method. The methods were tested in public databases and compared with the MLE method, obtaining superior results in classification problems.

Xue and Titterington (2010) propose (JoDiG) Joint Discriminative Generative Modelling. The method performs the parameters learning by dividing the variables into two sets: Discriminative and generative. A variable is treated in a discriminatory manner if the process or function that originates the data is not found, that is, if it does not have a good adherence to some probability distribution function. The method was tested on public databases of the UCI repository and obtained similar or better results than other methods that are only discriminative or generative.

Jing *et al*. (2011) propose a method of parameters learning based on the theory of interactive control of learning. The proposed algorithm provides the dynamic system and rules for upgrading CPT. The authors analyzed the convergence of the algorithm and concluded that the conditional probabilities reached reflected accurately to those desired. In addition, the convergence rate has been significantly improved when compared to other learning algorithms in the literature.

Carvalho *et al*. (2011) propose a data-based score metric without the use of parameters through Conditional Log-Likelihood factoring (CLL). The technique is used both for the structure learning and for parameters learning aiming to increase the classification in BN. The authors obtained good results by comparing the proposed method with other classifiers considered state of the art on public databases. In addition, the authors concluded that the computational time of the technique is significantly lower.

Broeck *et al*. (2014) propose a new family of algorithms for parameters learning considering missing data. The main features are: Parameters are computed in a non-interactive way, estimates are obtained without the need for Bayesian inference and the estimation of parameters is consistent for large databases. The authors conclude that the algorithms are faster than EM and avoid local minima.

This paper aims to explore the hybridism between fuzzy, basic statistics and Bayesian inference to compose an online method of parameters learning that is able to combine elements of generative and discriminative learning in Bayesian Networks.

Bayesian Networks

A Bayesian Network (BN) Pearl (1988) is a model of representation and reasoning of uncertainty that uses the conditional probability between variables of a specific domain, expressed by Directed Acyclic Graphs (DAG). Its graphical structure can tackle correlations between variables effectively, with appropriate language and efficient resources to represent the joint probability distribution over a set of random variables (Friedman and Goldszmidt (1996).

Defining formally, a BN is a pair (S, P), where S = (X, E) is a DAG. The nodes $X = \{X_1,\ldots,X_n\}$ represent the variables and edges $E = \{e_1,\ldots, e_m\}$ represent a direct correlation between each node in X.

P is defined as a set of probabilistic parameters expressed through tables. Given a particular variable, a conditional probability distribution is made for each of their classes/values $X_i = \{x_i^1,\ldots,x_i^k\}$ joining each classes/value of their parents Pa_i.

With that configuration, the network establishes that a variable is independent of all other variables except their descendants in the graph, given the state of its parents. The inference inside the network is done by the Bayes theorem for $P\left(X_i = x_i^k | pa_i = pa_i^j\right)$.

The joint probability is determined by the called chain rule and assumes the conditional independence between the variables:

$$P\left(X_1, X_2, \ldots, X_n\right) = \prod_{i=1}^{n} P\left(X_i | pa_i\right) \quad (1)$$

where, Pa_i determines the set of parent nodes from X_i.

The BN reasoning is established in two distinct scenarios:

$$\begin{cases} \text{if “input” then “output”} \\ \text{if “output” then “input”} \end{cases}$$

Considering all the possible network topologies for a Bayesian network the well-known structure Na¨ıve Bayes is the simplest one. It assumes that all variables are mutually independent given the class context. Although this model does not reflect the reality in most real-world tasks it is very effective, because the parameters of each attribute can be learned separately, facilitating the learning process McCallum and Nigam (1998). The na¨ıve

Bayes topology is there for a set of mutually independent variables that works as the input which collectively has a single parent (output node).

Parameters Learning

Parameter learning is related to filling the CPT in a fixed structure *S**. That is, it is assumed that there is a joint mdistribution of probability *P* (.) that represents a domain.

Generative Parameters Learning

Generative learning is made from the data set, seeking the maximization of likelihood Zhou (2015) and is known as (MLE) Maximum Likelihood Estimation) Friedman *et al.* (1997). The MLE estimate for each CPT after *T* samples, without missing data, is given by the formula:

$$\theta_{ijk}^{T} = \frac{N_{ijk}^{T}}{N_{ij}^{T}} \tag{2}$$

where, N_{ijk}^{T} is the number of times that the data was observed in the configuration x_i^k for the parent set pa_i^j and N_{ij}^{T} the total amount of X_i.

Parameters Learning with Missing data

Missing data can be divided into three categories Rubin (1976):

- MCAR: Missing completely at random
- MAR: Missing at random
- NMAR: Missing not at random

Missing data of type MCAR are those that have the highest degree of randomness and occur when the likelihood of finding a missing value is the same for all variables in any dataset. For example: In a network of sensors some of them, randomly, fail to capture data at certain times.

Data of type MAR occur when a variable X_j of the dataset influences the existence of missing data in a different variable X_i. For example, imagine a network of security sensors that capture the temperature and the existence of movement in a particular environment. Also imagine that some motion sensors have environmental-sensitive hardware: In the case of higher temperatures they cannot always capture the existence of movement. In this case a variable other than the one observed changes the likelihood of missing data happening.

Missing data is considered as NMAR when they are related to unobserved events or even the attribute itself. For example, if the ambient temperature influences the ability of the sensor to capture the data of the temperature itself or even if the factor influencing the occurrence of missing data is unknown.

Parameters learning with missing data can be summed up in three different approaches:

- Ignore/Discard data: It is the simplest way to deal with missing data, because it removes a data entry or even a variable. It is not always feasible and can generate large data distortions and is only recommended in MCAR cases
- Imputation: Technique that replaces the missing values with estimated values. The estimate may be by statistical measures obtained by the data or by some other technique of artificial intelligence. A good summary of the subject is found at Silva (2010)
- Parameters estimation: Methods that use the likelihood in the estimation. Two techniques are generally used: (EM) Expectation Maximization or likelihood optimization with a gradient-type method and are known to consistently estimate data of type MAR Broeck *et al.* (2014)

A variable with missing data is not a variable of hidden type: So there is data from the variable, but not in all cases.

The EM algorithm Dempster *et al.* (1977) enables the parameters estimation in models with missing data and is the most used algorithm in literature Zhou (2015). This algorithm uses a reiteration system that toggles in two steps (E step and M step) until it reaches convergence.

In a given instance y_l it is possible to have missing data ($Z_l = \{z_{l1},\ldots, z_{lo}\}$) and observed variables ($\Gamma_l = \{\gamma_{l1},\ldots, \gamma_{lh}\}$) where $o + h = n$. The steps for convergence are given by:

- E Step (expectation step): From the current parameters setting ($\theta^{(t)}$), where the first interaction is given by $\theta^{(0)}$ and has the initial configuration given by random values. Expectation is calculated through the maximum likelihood function considering the data set D:

$$l\left(\theta\middle|\theta^{(t)}\right) = \sum_{l} \sum_{\gamma_{l1},\ldots,\gamma_{lh}} P\left(\gamma_l\right) \log P\left(z_l\right) \tag{3}$$

where, $P\left(\gamma_l\right) = P\left(\gamma_{l1},\ldots,\gamma_{lh}\middle| Z_l,\theta^{(t)}\right)$ and $P\left(Z_l\right) = P\left(Z_l,\gamma_{l1},\ldots,\gamma_{lh}|\theta^{(t)}\right.$.

- (M Step) Maximization Step: Calculates the new estimation of $\theta^{(t+1)}$ parameters by maximizing the first step:

$$\theta^{(t)+1} = \arg\max_{\theta} l\left(\theta / \theta^{(t)}\right) \tag{4}$$

The Algorithm 1 describes the computational approach of EM

Algorithm 1 *Expectation Maximization* (EM)

```
θ ← random values
while not converge do
    Step E: use $\gamma_l$ to calculate $l(\theta|\theta^{(t)})$
    Step M: replace θ by arg max$_\theta$ $l(\theta|\theta^{(t)})$
end while
return θ
```

Another type of learning approach in missing data uses gradient methods, which are an alternative to learning in cases where BN has continuous variables Binder *et al.* (1997); Buntine (1994).

Other forms of learning in missing data were developed in the literature whether using methods of Monte Carlo or even by Gaussian approximation Barber (2012). In addition there are mixed approaches, such as that of Johnny that proposes a method of learning with focus on data of type MCAR and MAR through a BN that represents the relationship between these variables Mohan *et al.* (2013).

Discriminative Learning of Parameters

Discriminative learning is characterized when the main objective is the increase of accuracy in BN. However, discriminative learning has a computational complexity greater than generative and is considered an NP-hard problem *NP-hard* Greiner and Zhou (2002).

In this type of learning the goal is to find parameters that maximize the conditional log-likelihood as opposed to simply maximizing the likelihood. However there is no closed formula to find the best parameters of the network, since the conditional likelihood cannot be decomposed Friedman *et al.* (1997). One of the consequences of this is discriminative learning to generally use heuristic search methods to establish conditional probabilities Su *et al.* (2008). Or hybrid approaches between discriminative algorithms and generative as in Raina *et al.* (2003); Xue and Titterington (2010); Kang and Tian (2006).

Among the surveys in this area, it is possible to quote those with a purely discriminatory approach (Greiner and Zhou (2002); Greiner and Zhou (2002); Pernkopf and Wohlmayr (2009); Pernkopf and Bilmes (2005); Zhang and Su (2008); Carvalho *et al.* (2011); Feelders and Ivanovs (2006); Su *et al.* (2008).

Online Parameter Learning

In machine learning, online learning methods are those that learn from a set of data available in a sequential or interactive way. It is a type of adaptive learning and considers that the domain changes with time: The opposite of learning by batch, in which all data is available at the time of training.

Some of the algorithms most commonly used in the BN context use generative learning, such as in Cohen *et al.* (2001b) that proposes the VotingEM method based on the rules defined by Bauer *et al.* (1997) using concepts of maximum likelihood.

VotingEM

The Voting EM algorithm in Cohen *et al.* (2001b) is a direct adaptation of the EM (η) to be used online. The update rule is given by:

$$\theta_{ijk}^{t} = \begin{cases} \theta_{ijk}^{t-1} + \eta \left[\dfrac{p\left(x_i^k, pa_i^j \middle| y_t, \theta_{t-1}\right)}{p\left(pa_i^j \middle| y_t, \theta_{t-1}\right)} - \theta_{ijk}^{t-1} \right] & if\ P\left(pa_i^j \middle| d_t\right) \neq 0 \\ \theta_{ijk}^{t-1} & \text{otherwise} \end{cases} \tag{5}$$

where, d_t = (y_t, θ_{t-1}), T = {0,… t,…} is the current temporal unit and θ_{ijk}^{0} is populated by random or pre-trained values.

The learning rate η shows how much the past is reliable considering the data present. When η approaches 1 we consider the present data more reliable and the past knowledge is gradually discarded. The rate can be fixed for all learning or change over time (Section 3.4.2).

Adaptive VotingEM

One of the critical points of Voting EM is determining the learning rate η, because the parameter choice varies according to the application domain. In addition, a specific case x_j^k with parent configuration pa_i^j can be very constant or rarely appear in the database. With a fixed ETA the data influence on the CPT is always the same for all variables which makes the algorithm generic.

As a way to deal with the problem Cohen *et al.* (2001a) proposes the Adaptive Voting EM. It is based on the following principles:

- The learning rate η should be reduced when approaching convergence
- η should be increased when there is a large error between the average values of θ_{ijk} e θ_{ijk}^{t}
- A value η is defined for each pa_i^j being named η_{ij}

The method is based on traditional VotingEM, but the eta_{ij} value is updated on each time interaction and it uses 3 parameters as input:

- *q*: parameter that defines how many standard deviations of error is acceptable before increasing η_{ij}
- α: parameter that defines what is considered convergence in order to decrease η_{ij}
- *m*: parameter that defines in which proportion η_{ij} will be increased or decreased

The method variance is calculated by:

$$\operatorname{var}\left[\theta_{ijk}^{t}\right] = \frac{\eta_{ij}.0.5(1-0.5)}{2-\eta_{ij}} * \left(1-\left(1-\eta_{ij}\right)^{2\delta t+2}\right) \tag{6}$$

Cohen *et al*. (2001a) proves that η_{ij} decreases proportionally to $1/t_n$ where t_n is the number of times that $Pa_i = pa_i^j$ which leads to an optimal asymptotic convergence at some local maximum.

Our Proposal

In this work we propose a hybrid method (discriminative and generative) that addresses incremental or online learning of parameters in Bayesian networks through a fuzzy system. This method is based on the VotingEM algorithm Cohen *et al*. (2001b) that proposes an incremental version of the EM (η) Bauer *et al*. (1997).

Figure 1 synthesizes the proposed fuzzy system that has two types of input variables: Trend and classification error. The output variable is the adjustment level m that determines the variation of the learning rate η_{ij}. The method is described on Algorithm 2.

Algorithm 2 Our approach for online learning

```
$\theta^0 \leftarrow$ random or pre-trained values
$\eta_{ij} \leftarrow \eta' \mid 0 < \eta' \leq 1$ (randomly defined)
$t \leftarrow 0$
$\square_t \leftarrow 0$
while new samples do
    for all $X_i$ in $X$ do
        gets the set $Pa_i$
        for all $pa_i^j$ in $Pa_i$ do
            hypothesis = []
            for all $x_i^k$ in $X_i$ do
                update $\theta_{ijk}^{t+1}$
                $P_{cs} \leftarrow$ trend on $x_i^k$
                $hx_i^k$ hypothesis (fuzzy system)
                append $hx_i^k$ on hypothesis vector
            end for
            error $\leftarrow$ Classification Error
            m $\leftarrow$ fuzzy result (error, hypothesis)
            $\eta_{ij} \leftarrow \eta_{ij} . m$
            if $m > 1$ then
                $\delta t \leftarrow 0$
            else
                $\delta t \leftarrow 1$
            end if
        end for
    end for
$t \leftarrow t + 1$
end while
```

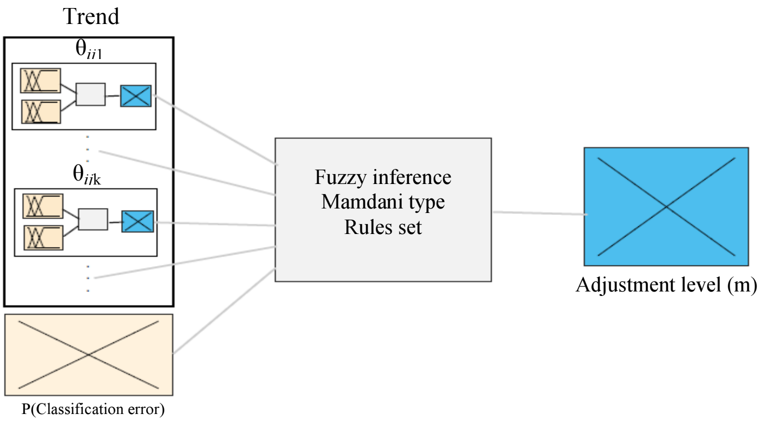


Fig. 1: Proposed fuzzy system

Definition of Variables

The following definitions are made:

- **Trend:** Long-term behavior of the time series that can be constant, growth or degrowth. Here, the term trend will be adopted in cases of growth or degrowth
- **Convergence:** Constant trend behavior-when the series focuses around a certain point
- **Precision:** Variation degree of a set of measures, in this case the variation of θ_{ij}. A significant accuracy in θ_{ij} demonstrates its convergence
- **Accuracy:** Network hit in a classification problem. In the discriminative approach it is portrayed by the a posteriori conditional probability of the output variable
- **Classification error:** Is the complementary measure to the conditional probability class_error = 1-*P* (.|.)

Trend

In online learning, the data $D = \{y_1,\ldots,y_N\}$ is not fully available during network training and is made available incrementally in time, featuring a temporal series.

Online learning methods are usually approached in a generative way and determine the influence of a evidence y_t in the CPT set θ that defines BN. The purpose of these methods is to reproduce the distribution of data in θ and the convergence is achieved by decreasing the influence of y_t in a consecutive way in $1/t$ Cohen *et al.* (2001b).

The influence of the learning rate is achieved by η_{ij} and the proposed method determines its value to each interaction of time. Similarly to VotingEM, the η_{ij} rate is increased when the error in θ_{ij} is considered relevant and diminished when θ_{ij} is converging. The difference between the VotingEM and the proposed method is in the approach and the concept of error and convergence.

The proposed method is based on the concept of tendency: Considering that *D* is a time series we define qt as the CPT set that compose the BN in time *t*. Determine whether a series has a tendency is usually carried out through statistical tests at a level of α significance, with two assumptions:

$$\begin{cases} H_0: & \text{The data is independent and identically distributed} \\ H_1: & \text{The data have monotonic tendency in time.} \end{cases}$$

In this work are used two tests of the literature to determine the trend of the series: Mann-Kendall and Cox-Stuart tests. The tests assess whether or not there is a tendency by calculating the p-value and comparing it with α.

The trend tests are performed for each θ_{ijk}, that is, for each pa_{ij} set of X_i. However, since the value of the learning rate is defined for η_{ij} the trend should be calculated for the whole set $\theta_{ij} = \{\theta_{ij1},\ldots\ \theta_{ijk},\ldots\}$ in a global way.

In addition to determining whether or not there is a tendency to θ_{ij}, other issues are raised:

- How to quantify the trend?
- What would be a statistically significant trend? And a statistically non-significant trend?
- How to use the quantification of the trend as a precision measure for θ_{ij}

In this work, a Fuzzy subsystem proposed using the p-values of Mann-Kendall and Cox-Stuart tests in order to answer these questions. The use of fuzzy functions to represent a statistical test was based on Costa (1999) that divides the p-value into three fuzzy sets: Highly significant, significant and non-significant (Fig. 2). The Trend Fuzzy subsystem is shown in Fig. 3.

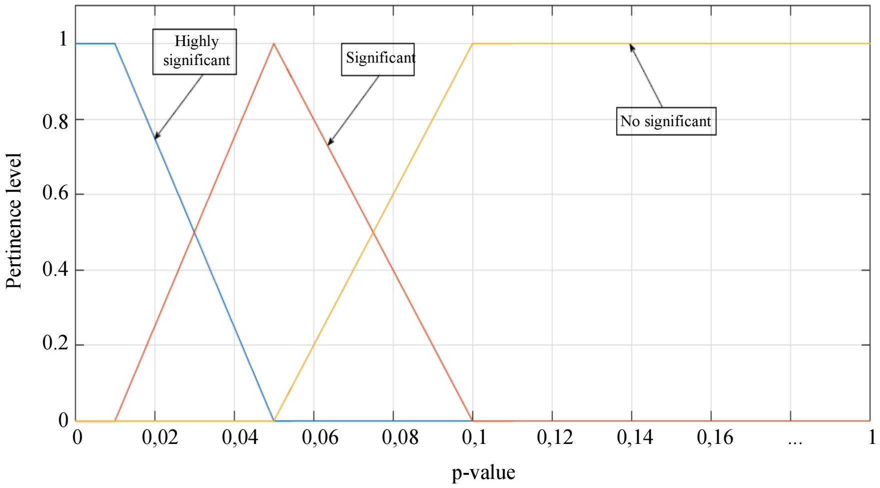


Fig. 2: Fuzzy sets for p-value

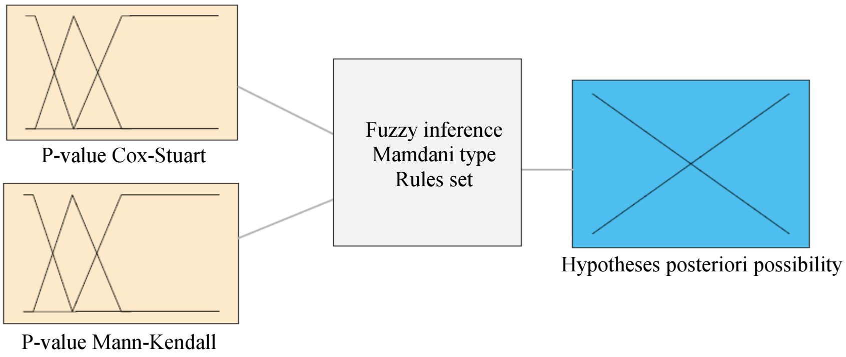


Fig. 3: Trend fuzzy subsystem for trend detection in θ_{ij}

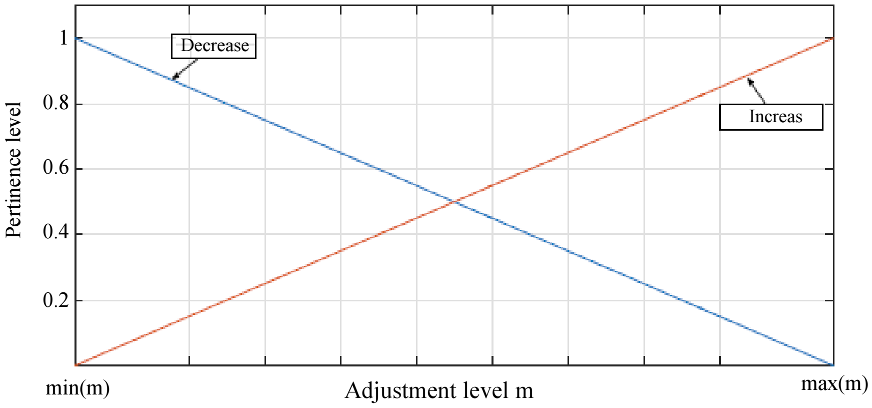


Fig. 4: Fuzzy sets for the adjustment level *m*

Trend tests are not applied to the entire history of θ_{ij}, but to moving temporal windows. Is defined as $\Theta_{ij}^{T}=\{\theta_{ij}^{1},...,\theta_{ij}^{1},...\}$ as the temporal window such as Θ_{ij}^{T}. A new temporal window Θ_{ij}^{T+1} is created whenever the ηij value is increased, that is when convergence in the series is not detected.

Classification Error

The proposed method, in addition to seeking the precision in θ_{ij} seeks to increase accuracy of the network considering an output variable X_s. That is, seeks to decrease the classification error, calculated by:

$$P(class\_error)=P\left(x_s^k\,|\,\rho_{ij},\theta_{t-1}\right)-P\left(x_s^k\,|\,\tau_t,\theta_{t-1}\right) \tag{7}$$

where, X_s is the output node in a classification problem that assumes a value x_s^k at time t, ρ_t is the set with the evidence of X_s at time t and τ_t are the evidence of the input data, such that $y_t=\rho_t\cup\tau_t\rho_t\cap\tau_t=\emptyset$. When there is no missing data we can reduce the Equation 7 to:

$$P(class\_error)=1-P\left(x_s^k\,|\,\tau_t,\theta_{t-1}\right) \tag{8}$$

Fuzzy Inference

The change in the learning rate η_{ij} is made by the combination between the error rate of X_s and the quantification of the trend in θ_{ij}. This combination of factors is made by a fuzzy system that has a set of input variables and an output variable called the adjustment level *m*, such that:

$$\eta_{ij}=\eta_{ij}.m \tag{9}$$

The inputs in the proposed subsystem are defined by:

- The output variables of the Fuzzy Trend Subsystem shown in the Fig. 3 for every θ_{ijk} in θ_{ij} in the moving window Θ_{ij}^{T}
- The error of the output variable X_s in time t (Equation 7)

Adjustment level m

The value of *m* is obtained by a fuzzy inference system of Mandeni type and defined by two fuzzy sets that model its behavior by increasing or diminishing η_{ij} (Fig. 4).

Preliminary Results

The initial evaluation of the method performance was done by simulating an online environment considering the Bayesian network provided by Cohen *et al*. (2001b).

The proposed method was compared with two other methods of literature: VotingEM and MLE Online. The choice of these methods for initial comparison is due to its great popularity, efficiency and low computational cost Liu and Liao (2008; Zhou, 2015).

The initial experiments seek to evaluate the method in two distinct conditions:

- **Condition** 1: Considering a totally random BN
- **Condition 2:** Considering a pre-trained BN that has undergone abrupt changes in its probability distribution

Condition 1 is simulated through 2000 independent and equally distributed (i.i.d.) samples generated from the CPT. To simulate learning, a random BN is created and the samples generated from the original BN are sent interactively to learning methods. The resulting BN after the simulation of Condition 1 is used as input in Condition 2.

Condition 2 is simulated by the abrupt change in the distribution of conditional probability of some values in θ_{ij} at the network obtained at the end of Condition 1. After the change, in a similar way to Condition 1, 2000 i.i.d. samples were generated from this new BN. The objective of Condition 2 is to analyze the capacity of the proposed method of adapting to changes in the environment.

The results evaluation is done considering Condition 1 and Condition 2. The Bayesian network (BN) used for the initial evaluation was proposed by Cohen *et al*. (2001b) and has three nodes: Parent, Child1 and Child 2 (adopted as gold standard). The CPT set that composes it is shown in Table 1. The gold standard BN was used for the generation of 2000 samples for Condition 1.

Condition 2 is simulated by changing three θ_{ij} values in the gold standard BN: one in each X_i node. Similarly to condition 1, 2000 samples were generated in the altered network. The experiment was performed using the parameters defined in Table 2 obtained empirically through experimentation. The MLE Online method does not require an initial parameterization.

Figure 1, 5 shows convergence of methods. Figures 5 (a) (b) and (c) demonstrate convergence in three θ_{ijk} parameters and the following observations are made:

- The proposed method and VotingEM have similar convergences in Condition 1
- The proposed method perceives changes in the environment faster (Condition 2)
- The variability of the proposed method is greater;
- By not enabling the increase in the rate of learning, the MLE Online method does not have a good convergence

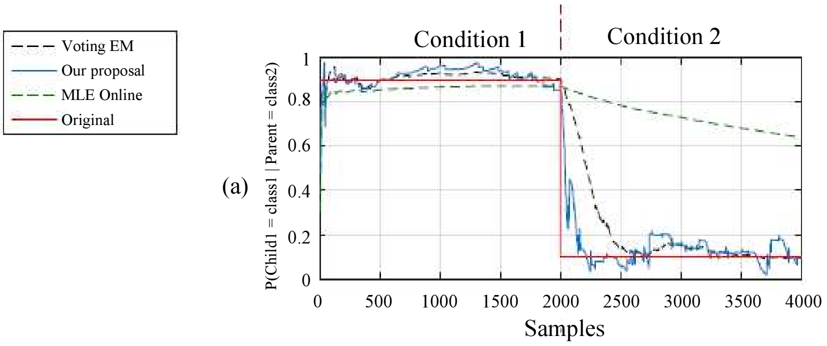
Figure 6 shows the overall convergence by log-likelihood of the trained BN at each new sample. The proposed method has a faster convergence in Condition 2 than the other two methods.

Table 1: CPT set for the Bayesian Network proposed by Cohen *et al*. (2001b)

Parent		
class1	0.5	
class2	0.25	
class3	0.15	
Child1	class1	class2
class1	0.5	0.5
class2	0.9	0.1
class3	0.85	0.15
Child2		
class1	0.8	0.2
class2	0.2	0.8
class3	0.85	0.15

Table 2: Parameters Configuration used during simulation of Condition 1 and Condition 2

Method	Parameter	Values
Voting EM	η inicial	0.25
	Q	4.00
	A	0.10
	m	1.50
Our proposal	η inicial	0.25
	max $\delta\eta$	0.02
	output node X_s	Parent



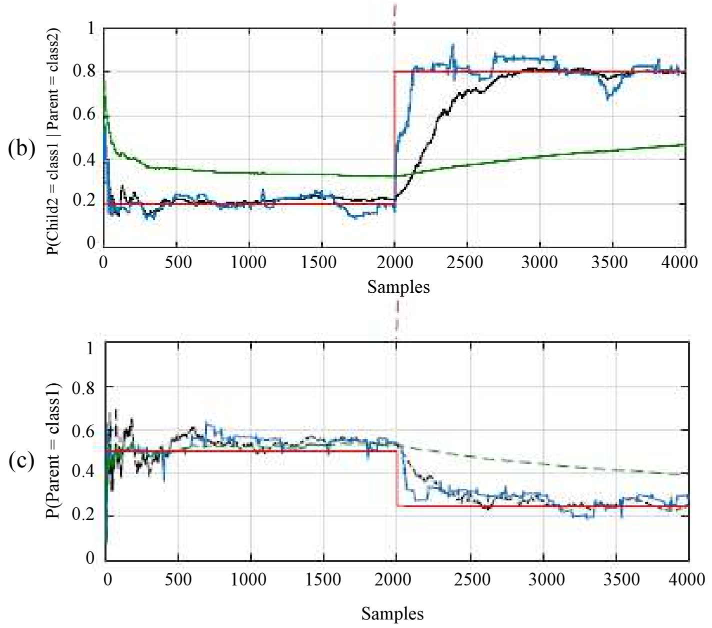


Fig. 5: Convergence of proposed method, voting EM and MLE Online. Figure 5(a) (b) and (c) show the evolution of three parameters θ_{ijk}

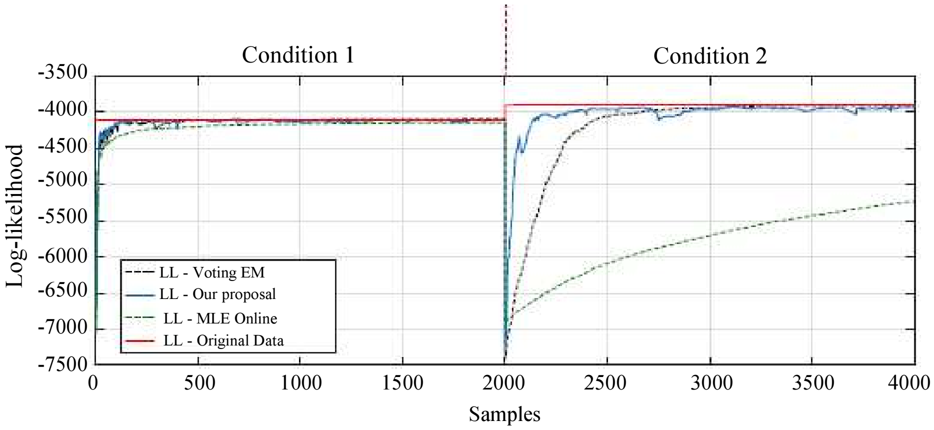


Fig. 6: Evolution of the BN log-likelihood in condition 1 and condition 2

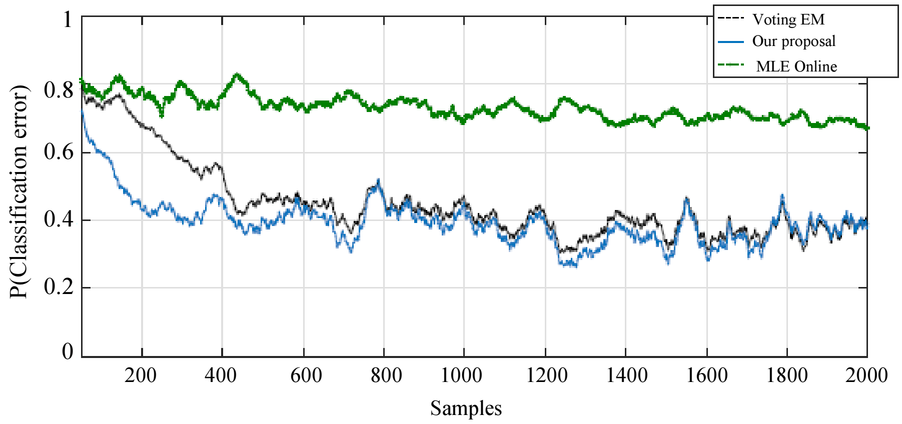


Fig. 7: Moving averages with 50 sample demonstrating the evolution of the classification error in condition 2

The discriminative aspect is analyzed by Fig. 7, which reinforces the ability to decrease the error probability in the proposed method and its rapid convergence to lower errors. Figure 1 also demonstrates the abrupt drop in error by using the proposed approach reaching stability in sample 400, approximately. The VotingEM method only arrives at this stability around the sample 600.

Conclusion

The initial analysis of the results shows that the proposed method achieved a good performance both in the convergence η_{ijk} and in the decrease of the probability of the classification error when compared to the other methods.

The proposed approach also simplifies parameterization by using only one configuration parameter while VotingEM uses three.

The following observations can be made:

- The proposed method is more sensitive to the environment, which results in a greater variability in the estimation of θ_{ijk}
- MLE Online is not able to increase the learning rate during the simulation
- The proposed method perceives the distribution change faster than VotingEM and increases η_{ijk} more significantly

The results reinforce the main characteristics of the proposed method: Perceive changes of distribution rapidly and unite the generative and discriminative approaches during learning. For this reason, there is a greater variation of probability distributions as a resource to decrease the probability of sort error.

Acknowledgement

This study was financed in part by the Coordenação de Aperfeiçoamento de Pessoal de Nível Superior - Brasil (CAPES) - Finance Code 001.

Author's Contributions

Both authors contributed equally to this research.

Ethics

This study is self-contained and includes unpublished material. The authors confirm that they have read and approved this document and there is no ethical issue involved.

References

Barber, D., 2012. Bayesian Reasoning and Machine Learning. 1st Edn., Cambridge University Press. ISBN-10: 0521518148, pp: 697.

Bauer, E., D. Koller and Y. Singer, 1997. Update rules for parameter estimation in Bayesian networks. Proceedings of the 13th Conference on Uncertainty in Artificial Intelligence, Feb. 6, IEEE Xplore Press, San Francisco, CA, USA, pp: 3-13.

Binder, J., D. Koller, S. Russell and K. Kanazawa, 1997. Adaptive probabilistic networks with hidden variables. Machine Learning, 29: 213-244. DOI: 10.1023/A:1007421730016

Bridle, J.S., 1990. Training stochastic model recognition algorithms as networks can lead to maximum mutual information estimation of parameters. Adv. Neural Inform. Process. Syst.

Brignoli, J.T., 2013. Um Modelo para Suporte ao Raciocinio Diagnostico diante da Di-namica do Conhecimento sobre incertezas. Thesis PhD, Universidade Federal de Santa Catarina.

Broeck, G.V.D., K. Mohan, A. Choi and J. Pearl, 2014. Efficient algorithms for Bayesian network parameter learning from incomplete data. arXiv preprint arXiv:1411.7014.

Buntine, W.L., 1994. Operations for learning with graphical models. J. Artificial Intelli. Res., 2: 159-225. DOI: 10.1613/jair.62

Carvalho, A.M., T. Roos, A.L. Oliveira and P. Myllym¨aki, 2011. Discriminative learning of Bayesian networks via factorized conditional log-likelihood. J. Machine Learn. Res., 12: 2181-2210.

Chen, R., K. Sivakumar and H. Kargupta, 2001. An approach to online Bayesian learning from multiple data streams. Proceedings of the Workshop on Mobile and Distributed Data Mining, (KDD' 01), pp: 31-45.

Cohen, I., A. Bronstein and F.G. Cozman, 2001a. Adaptive online learning of Bayesian network parameters. University of Sao Paulo, Brasil.

Cohen, I., A. Bronstein and F.G. Cozman, 2001b. Online learning of bayesian network parameters. Hewlett Packard Laboratories Technical Report, HPL- 2001-55.

Costa, P.A.B., 1999. Um Enfoque Segundo a Teoria De Conjuntos Difusos Para a Meta-An´Alise.Thesis PhD, Universidade Federal de Santa Catarina.

Dempster, A.P., N.M. Laird and D.B. Rubin, 1977. Maximum likelihood from incomplete data via the em algorithm. J. Royal Stat. Soc. Series B (Stat. Methodol.), 39: 1-38.

Lima, M.D., Nassar, S.M., Freitas Filho, P.J. and Jacinto, C.M., 2014. Heuristic discretization method for bayesian networks.

Feelders, A. and J. Ivanovs, 2006. Discriminative scoring of Bayesian network classifiers: A comparative study. Proceedings of the 3rd European Workshop on Probabilistic Graphical Models, Sep. 12-15, pp: 75-82.

Friedman, N. and M. Goldszmidt, 1996. Discretizing continuous attributes while learning Bayesian networks. Proceedings of the 13th International Conference on International Conference on Machine Learning, Jul. 3-6, IEEE Xplore Press, Bari, Italy, pp: 157-165.

Friedman, N., D. Geiger and M. Goldszmidt, 1997. Bayesian network classifiers a comparative study. Machine Learning, 29: 131-163.

Greiner, R. and W. Zhou, 2002. Structural extension to logistic regression: Discriminative parameter learning of belief net classifiers.

Jing, C., F. Jing-qi and S. Wei, 2011. Learning Bayesian network parameters based on iterative learning control. Proceedings of the International Conference on Consumer Electronics, Communications and Networks (CECNet), Apr. 16-18, IEEE Xplore Press, XianNing, China, pp: 4161-4165. DOI: 10.1109/CECNET.2011.5768842

Kang, C. and J. Tian, 2006. A hybrid generative/discriminative Bayesian classifier. Proceedings of the 19th International Florida Artificial Intelligence Research Society Conference, May 11-13, Melbourne Beach, Florida, USA, pp: 562-567.

Kurihara, T., Y. Nakada, K. Yosui and T. Matsumoto, 2001. Bayesian on-line learning: A sequential monte carlo with importance resampling. Proceedings of the IEEE Signal Processing Society Workshop Neural Networks for Signal Processing XI, Dec, 12, IEEE Xplore Press, North Falmouth, MA, USA, pp: 163-172. DOI: 10.1109/NNSP.2001.943121

Liu, J. and Q. Liao, 2008. Online learning of bayesian network parameters. Proceedings of the 4th International Conference on Natural Computation, Oct. 18-20, IEEE Xplore Press, Jinan, China, pp: 267-271. DOI: 10.1109/ICNC.2008.651

McCallum, A. and K. Nigam, 1998. A comparison of event models for naive bayes text classification. Workshop, Learning Text Categorization, 752: 41-48.

Mohan, K., J. Pearl and J. Tian, 2013. Graphical Models for Inference with Missing Data. In: Advances in Neural Information Processing Systems, Burges, C.J.C., L. Bottou, M. Welling, Z. Ghahramani and K.Q. Weinberger, (Eds.), Los Angeles, Ames, pp: 1277-1285.

Myers, J.W., K.B. Laskey and K.A. DeJong, 1999. Learning Bayesian networks from incomplete data using evolutionary algorithms. Proceedings of the 1st Annual Conference on Genetic and Evolutionary Computation, Jul. 13-17, I EEE Xplore Press, Orlando, Florida, pp: 458-465.

Pearl, J., 1988. Probabilistic Reasoning in Intelligent Systems: Networks o f Plausible Inference. 2nd Edn., Morgan Kaufmann, ISBN-10: 0934613737, pp: 552.

Pernkopf, F. and J. Bilmes, 2005. Discriminative versus generative parameter and structure learning of Bayesian network classifiers. Proceedings of the 22nd International Conference on Machine Learning, Aug. 7, IEEE Xplore Press, Bonn, Germany, pp: 657-664. DOI: 10.1145/1102351.1102434

Pernkopf, F. and M. Wohlmayr, 2009. On Discriminative Parameter Learning of Bayesian Network Classifiers. In: Machine Learning and Knowledge Discovery in Databases, Buntine, W., M. Grobelnik, D. Mladenić and J. Shawe-Taylor, (Eds.), Springer, Berlin, Heidelberg, ISBN-10: 978-3-642-04173-0, pp: 221-237.

Raina, R., Y. Shen, A. Mccallum and A.Y. Ng, 2003. Classification with hybrid generative/discriminative models. Proceedings of the 16th International Conference on Neural Information Processing Systems, Dec. 09-11, IEEE Xplore Press, Whistler, British Columbia, Canada, pp: 545-552.

Ratnapinda, P. and M.J. Druzdzel, 2015. Learning discrete Bayesian network parameters from continuous data streams: What is the best strategy? J. Applied Logic. 13: 628-642. DOI: 10.1016/j.jal.2015.03.007

Rubin, D.B., 1976. Inference and missing data. Biometrika, 63: 581-592. DOI: 10.2307/2335739

Saloj¨arvi, J., K. Puolam¨aki and S. Kaski, 2005. Expectation maximization algorithms for conditional likelihoods. Proceedings of the 22nd International Conference on Machine Learning, Aug. 07-11, IEEE Xplore Press, Bonn, Germany, pp: 752-759. DOI: 10.1145/1102351.1102446

Silva, J.D.A., 2010. Substituição de valores ausentes: uma abordagem baseada em um algoritmo evolutivo para agrupamento de dados. Thesis PhD, Universidade de S˜ao Paulo.

Su, J., H. Zhang, C.X. Ling and S. Matwin, 2008. Discriminative parameter learning for Bayesian networks. Proceedings of the 25th International Conference on Machine Learning, Jul. 05-09, IEEE Xplore Press, Helsinki, Finland, pp: 1016-1023. DOI: 10.1145/1390156.1390284

Xue, J.H. and D.M. Titterington, 2010. Joint discriminative–generative modeling based on statistical tests for classification. Pattern Recognition Lett., 31: 1048-1055.

Zhang, H. and J. Su, 2008. Naive bayes for optimal ranking. J. Experimental Theoretical Artificial Intelligence, 20: 79-93.

Zhang, S.Z. and L. Liu, 2008. Mcmc samples selecting for online Bayesian network structure learning. Proceedings of the International Conference on Machine Learning and Cybernetics, Jul. 12-15, IEEE Xplore Press, Kunming, China, pp: 1762-1767. DOI: 10.1109/ICMLC.2008.4620690

Zhou, Y., 2015. New techniques for learning parameters in Bayesian networks. Thesis PhD, Queen Mary University of London.