Stochastic and Physical Monitoring Systems
SPMS 2015
June 22 - June 27, 2015
Drhleny, Czech Republic

Lazy Learning of Environment Model from the Past

Jakub Štěch[1], Tatiana V. Guy[2], Barbora Pálková[1], Miroslav Kárný[2]

[1] Department of Mathematics,
Faculty of Nuclear Sciences and Physical Engineering,
Czech Technical University,
Trojanova 13, 120 00 Prague 2
Czech Republic

[2] Department of Adaptive Systems
Institute of Information Theory and Automation,
Czech Academy of Sciences,
P.O. Box 18, 182 08 Prague 8,
Czech Republic

Email: stech@seznam.cz

Abstract. The paper addresses a lazy learning (LL) approach to decision making (DM) problem described in fully probabilistic way. The key idea of LL is to simplify the actual DM problem by using past DM problems similar to the current one. The approach can decrease computation complexity and increase quality of learning when no rich alternative information available.

The proposed LL approach helps to learn the environment model based on a proximity of the past and current DM problem with Kullback-Leibler divergence serving as a proximity measure. The implemented algorithm is verified on the real data. The results show that the proposed approach improves prediction quality.

Key words: Lazy learning, local modelling, prediction for optimisation.

1 Introduction

To make a good decision with respect to an environment, a decision maker should have a good environment model and high-quality prediction of the environment behaviour, which can be used in optimisation. There are several approaches to make good environment model. In most learning methods, one global model is searched for to describe all of the past data. It is essential that every complex structure contains several subsystems. Each subsystem exhibits different behaviour, we need to identify each subsystem and analyse it separately.

This has motivated the approach, which builds local models using the relevant data from the past. This group of learning approaches, generally called lazy learning (LL), goes through the data history and search for the past instances similar to the current problem. Then the relevant past data is used to build a current environment model. These models attempt to fit the past data only in a region around the location of the query. The strong motivation for local learning techniques is their suitability for real time tasks. Their fast learning and avoidance of negative interference between historical and new data is a great advantage. The article addresses employing LL approach to improving the quality of the prediction further used in optimisation. The resulting solution was applied to real data.

2 Preliminaries

This section introduces necessary notions and definitions. The sequence $(x_t, x_{t-1}, \ldots, x_1)$ is shortened as $x(t)$. Values of x given in discrete time instances are labelled by t, $t \in \mathbb{N}$. If x is a vector, x' is its transpose and l_x its length. Bold capital $\mathbf{X}$ represents a set of x values. If x is a random variable, ${}^c x$ is a realisation of x. We use an abbreviation *pdf* for probability density function.

Consider an interacting pair 'environment-agent'. The agent observes environment state $s_t \in \mathbf{S}$ and makes action $a_t \in \mathbf{A}$ to learn (or influence) the environment. The selected actions are expected to provide the desired behaviour of the pair. The complete probabilistic description of the closed-loop behaviour of the pair up to time t is represented by a joint pdf $p(s(t), a(t))$, that can be decomposed using the chain rule for pdfs, [5],

$$p(s(t), a(t)) = \prod_{\tau=1}^{t} p(s_\tau | a_\tau, s(\tau-1), a(\tau-1)) p(a_\tau | s(\tau-1), a(\tau-1)) p(s_0), \qquad (1)$$

where $p(s_\tau | a_\tau, s(\tau-1), a(\tau-1))$ is a model of the environment, $p(a_\tau | s(\tau-1), a(\tau-1))$ stands for the agent's decision rule forming DM policy and $p(s_0)$ is the prior pdf.

2.1 Environment Model

We consider the following model of the environment.

Definition 1 *(Model of the environment)*
Let a time-invariant parametrised environment model $M(\Psi_t, \Theta)$ specify a pdf of the environment observed state s_t given at most $N \in \mathbb{N}$ past data:

$$\begin{aligned} M(\Psi_t, \Theta) &= p(s_t | s_{t-1}, \ldots, s_{t-N}, a_t, \ldots, a_{t-N}, \Theta) = p(s_t | \psi_t, \Theta) \\ \Psi_t &= [s_t, \ldots, s_{t-N}, a_t, \ldots, a_{t-N}]' = [s_t, \psi_t']', \end{aligned} \qquad (2)$$

where $\psi_t = [s_{t-1}, \ldots, s_{t-N}, a_t, \ldots, a_{t-N}]'$ is a regression vector, Ψ_t is a data vector and $\Theta \in \mathbf{\Theta}$ is an unknown finite-dimensional parameter.

The regression vector ψ_t consists of a fixed number of past (delayed) actions and states. $N \in \mathbb{N}$ is a maximal delay of past data entering the regression vector. The data

vector $\Psi_t \in \mathbf{\Psi}$ is recursively updated, $(\Psi_t, s_{t+1}, a_{t+1}) \implies \Psi_{t+1}$ when new data come. Knowledge of the unknown parameter $\Theta \in \mathbf{\Theta}$ is described by a flat prior pdf $p(\Theta)$.

Bayesian estimation [3] is used for modifying the prior pdf $p(\Theta)$ to the posterior pdf $p(\Theta|\mathcal{K})$. The resulting $p(\Theta|\mathcal{K})$ reflects the knowledge $\mathcal{K}$ gained from the past data vectors and is further used to get a prediction $p(s_t|\psi_t, \mathcal{K})$ serving as the environment model

$$p(\Theta|\mathcal{K}) \propto \prod_t p(\Theta) M(\Psi_t, \Theta), \tag{3}$$

$$p(s_t|\psi_t, \mathcal{K}) = \int_\Theta M([s_t', \psi_t']', \Theta) p(\Theta|\mathcal{K}) d\Theta, \tag{4}$$

where $\propto$ is the proportional sign.

2.2 Estimation of Structure and Parameters of Linear Normal Environment Model

The considered application deals with linear normal model $M(\Psi, \Theta) = \mathcal{N}_s(\theta'\psi, r)$ and

$$s_t = \theta'\psi_t + e_t, \tag{5}$$

where θ is unknown time-invariant matrix of regression coefficients $(l_\psi \times l_s)$, ψ_t is a regression vector of length $l_\psi = 2N + 1$ and e_t is normally distributed random variable (noise), i.e.

$$p(e_t|a(t), s(t-1), r) = p(e_t|r) = (2\pi)^{-l_s/2} |r|^{-1/2} \exp\left\{-\frac{1}{2} e_t' r^{-1} e_t\right\}, \tag{6}$$

where r is positive-definite $(l_s \times l_s)$ covariance matrix. In order to get the model (4) we need to evaluate $p(\Theta|\mathcal{K})$, (3), where $\Theta = (\theta, r)$.

Let us denote $\hat{s}_t = \theta'\psi_t$ and assume the structure of the model (5) to be fixed within a hypothesis denoted H_ψ. Using (6) and the fact that $e_t = s_t - \hat{s}_t$ we can calculate

$$p(s_t|a(t), s(t-1), \Theta, H_\psi) = (2\pi)^{-l_s/2} |r|^{-1/2} \exp\left(\mathrm{tr}\left(r^{-1} \begin{bmatrix} I_{l_s} \\ -\theta \end{bmatrix}' \begin{bmatrix} s_t \\ \psi_t \end{bmatrix} \begin{bmatrix} s_t \\ \psi_t \end{bmatrix}' \begin{bmatrix} I_{l_s} \\ -\theta \end{bmatrix}\right)\right), \tag{7}$$

where I_k is $(k \times k)$ unit matrix. We assume the class of input generators satisfying the natural control condition [3], i.e.

$$p(a_t|a(t-1), s(t-1), \Theta, H_\psi) = p(a_t|a(t-1), s(t-1)). \tag{8}$$

It can be shown, using (7), that pdf of an observed part of the environment state takes the form

$$p(s(t), a(t)|H_z) \propto \prod_{\tau=1}^{t} p(a_\tau|a(\tau-1), s(\tau-1)) p(a_0, s_0|H_\psi) \left|\frac{V_{\psi t}}{\varepsilon}\right|^{-l_s/2} |\Lambda_t|^{-\nu_t/2} \tag{9}$$

where

$$V_t = V_{t-1} + \begin{bmatrix} s_t \\ \psi_t \end{bmatrix} \begin{bmatrix} s_t' & \psi_t' \end{bmatrix} = \begin{bmatrix} V_{st} & V_{\psi st}' \\ V_{\psi st} & V_{\psi t} \end{bmatrix}, \; V_0 = \varepsilon I_{l_\Psi}$$

$$\Lambda_t = V_{st} - V_{\psi st}' V_{\psi t}^{-1} V_{\psi st}, \; \nu_t = \nu_{t-1} + 1, \; \nu_0 = \varepsilon$$

and $\varepsilon > 0, \varepsilon$ small (standard choice in case of insufficient prior information, see [6]).

A finite amount of possible structures of model given by $\psi^i, i \in \{1, 2, \ldots, M\}$, is assumed. The probabilities of hypotheses $H_i = H_{\psi^i}$ are calculated according to the Bayes rule

$$p(H_i|s(t), a(t)) = \frac{p(s(t), a(t)|H_i)p(H_i)}{\sum_{k=1}^{M} p(s(t), a(t)|H_k)p(H_k)}. \tag{10}$$

We have no reason to expect any hypothesis to be more probable than others, hence $p(H_i) = \frac{1}{M}$. This, use of (10) and natural conditions of control (8) implies [6]

$$p(H_i|s(t), a(t)) = \frac{\gamma_{i(t)}}{\sum_{k=1}^{M} \gamma_{k(t)}}, \quad \gamma_{i(t)} = |V_{\psi^{i(t)}}|^{-l_s/2} |\Lambda_{i(t)}|^{-\nu_t/2} \varepsilon^{l_\psi l_s/2}. \tag{11}$$

We arrived to this using (9). A full description of pdf of observed data contains also factors describing the input generator and normalizing factors. However, this formula is simplified using (11), see [6].

The model is estimated to have a structure given by ψ^{ibest}, where

$$p(H_{ibest}|s(t), a(t)) = \max_{k \in \{1,2,\ldots,M\}} p(H_k|s(t), a(t))$$

In order to obtain the estimate of Θ, we have to determine for $\psi = \psi^{ibest}$ the pdf $p(\Theta|a(t), s(t), H_\psi) = p(r, \theta|a(t), s(t), H_\psi)$. In the linear normal case, we choose prior pdf in the Normal-inverse-Wishart form $\mathcal{N}i\mathcal{W}_{\theta,r}(V_0, \nu_0)$

$$p(\Theta|a_0, s_0, H_\psi) = p(\Theta|H_\psi) \propto |r|^{-\frac{\nu_t}{2}} \exp\left\{-\frac{1}{2}\mathrm{tr}\left(r^{-1} \begin{bmatrix} I_{l_s} \\ -\theta \end{bmatrix}' V_0 \begin{bmatrix} I_{l_s} \\ -\theta \end{bmatrix}\right)\right\}, \tag{12}$$

the pdf $p(r, \theta|a(t), s(t), H_\psi)$ can be rewritten (see [3]) like

$$\begin{aligned} p(\Theta|a(t), s(t), H_\psi) &\propto |r|^{-\frac{\nu_t}{2}} \exp\left\{-\frac{1}{2}\mathrm{tr}\left[r^{-1}\left(\left[\theta - \hat{\theta}\right]' V_t \left[\theta - \hat{\theta}\right] + \Lambda_t\right)\right]\right\}, \\ \hat{\theta}_t &= C_t V_{\psi st}, \quad C_t = V_{\psi t}^{-1}, \quad \Lambda_t = V_{\psi t} - V'_{\psi st} C_t V_{\psi st}. \end{aligned} \tag{13}$$

Note that maximum of the pdf is at $\theta = \hat{\theta}$. We can modify (13), see [3], to obtain recursions

$$\begin{aligned} \hat{\theta}_t &= \hat{\theta}_{t-1} + \frac{1}{1+\zeta_t} C_{t-1}\psi_t \hat{e}'_t, \quad C_t = C_{t-1} - \frac{1}{1+\zeta_t} C_{t-1}\psi_t \psi'_t C_{t-1} \\ \Lambda_t &= \Lambda_{t-1} + \frac{1}{1+\zeta_t} \hat{e}_t \hat{e}'_t, \quad \zeta_t = \psi'_t C_{t-1} \psi_t, \quad \hat{e}_t = s_t - \hat{\theta}'_{t-1}\psi_t. \end{aligned} \tag{14}$$

Recursions (14) are formally identical with the recursive least squares. Hence, this method is used to get estimate $\hat{\theta}_t$ of θ, [3].

3 Lazy Learning

Lazy learning is a learning method based on similarity of the actual and past data. Here we use the LL approach for selecting an environment model.

3.1 Formulation

Let us suppose that similar regression vectors indicate similarity of the actual state of the pair 'agent-environment' and some past state. Naturally this similarity yields similar prediction of the environment state and can be described by a single environment model. To decide on the proximity of the current state and the state occurred in the past, the Kullback-Leibler divergence (KLD) is used.

Definition 2 *(Kullback-Leibler divergence)*
The Kullback-Leibler divergence $\mathcal{D}(f||g)$ measures the proximity of two pdfs (f, g) acting on a set $\mathbf{X}$. It is defined as:

$$\mathcal{D}(f||g) = \int_{x\in\mathbf{X}} f(x) ln \frac{f(x)}{g(x)} dx.$$

It can be shown that $\mathcal{D}(f||g) \geq 0$, $\mathcal{D}(f||g) = 0$ iff $f = g$ almost everywhere on $\mathbf{X}$.

3.2 Lazy Learning of the Environment Model

At the very beginning a prior pdf $p(\Theta)$ of the unknown parameter is used for the Bayesian estimation. Consider the current time t (before observing state s_t). Let us have a set of knowledge-expressing past data vectors $\{{}^c\Psi_\kappa\}_{\kappa\in\kappa}$, $\kappa = 1, 2, \ldots, t-1$, which are used to update the prior to $p(\Theta|\mathcal{K})$, (3).

Let us have model $M(\Psi_t, \Theta)$. Its parameter is estimated at time t by applying the Bayes rule to a selected range of data indexed by κ, i.e. $\{{}^c\Psi_\kappa\}_{\kappa\in\kappa}$ (3), (4). At time t we can get the prediction of the environment state s_t

$$p(s_t|{}^c\psi_t, \mathcal{K}) = \int_\Theta M([s_t', {}^c\psi_t']', \Theta) p(\Theta|\mathcal{K}) d\Theta. \tag{15}$$

This whole design depends on the choice of the relevant (close) data vectors to be included into $\mathcal{K}$. Thus the definition of the closeness of regression vectors ${}^c\psi_t$ and ${}^c\psi_\kappa$ is very important. The similar regression vectors have similar joint pdf of the predicted state s_t and the parameter Θ, thus $\psi_t \approx \psi_\kappa$ implies

$$p(s_t, \Theta|{}^c\psi_t) \approx p(s_t, \Theta|{}^c\psi_\kappa).$$

No information about the parameter Θ can be gained from the regression vector (an analogy of natural conditions of control, [3]), i.e

$$p(\Theta|{}^c\psi_t) = p(\Theta|{}^c\psi_\kappa) = p(\Theta).$$

Hence, the Kullback-Leibler divergence (Definition 2) of $p(s_t, \Theta|{}^c\psi_t)$ on $p(s_t, \Theta|{}^c\psi_\kappa)$ reads

$$\mathcal{D}_{t\kappa} = \int_{\mathbf{s_t},\Theta} M([s_t', {}^c\psi_t']', \Theta) p(\Theta) \ln \frac{M([s_t', {}^c\psi_t']', \Theta)}{M([s_t', {}^c\psi_\kappa']', \Theta)} ds d\Theta.$$

For a single-output normal autoregressive model the considered conjugate pdf of the unknown parameter, Section 2.2, is Normal-inverse-Gamma (Wishart) distribution $\mathcal{N}i\mathcal{G}_{\theta,r}(V,\nu)$, (12), and KLD reads

$$\begin{aligned}\mathcal{D}_{t\kappa} &= \int_{\theta,r\geq 0} \frac{[\theta'({}^c\psi_t - {}^c\,\psi_\kappa)]^2}{2r}\mathcal{N}i\mathcal{G}_{\theta,r}(V,\nu)d\theta dr \\ &= \frac{1}{2}\Big[\frac{\nu}{(\nu-2)\hat{r}}[\hat{\theta}'({}^c\psi_t - {}^c\,\psi_\kappa)]^2 + ({}^c\psi_t - {}^c\,\psi_\kappa)'C^{-1}({}^c\psi_t - {}^c\,\psi_\kappa)\Big]. \end{aligned} \quad (16)$$

Here $\hat{\theta}$ is the least squares (LS) estimate of regression coefficients, $\Theta = (\theta, r)$, where θ consists of the regression coefficients of the environment model (5) and r is variance of the noise e_t, the number ν stands for degrees of freedom, $\hat{r}$ is a LS estimate of noise variance and C is a factor of LS covariance $\hat{r}C$. This result comes from the basic properties of Normal-inverse-Gamma pdf $\mathcal{N}i\mathcal{G}_{\theta,r}(V,\nu)$, see Section 2.2 and [1]. The values of KLD have to be small for similar regression vectors ${}^c\psi_t \approx^c \psi_\kappa$. The first summand in (16) is proportional to the normalised squared difference of the regression vectors based on estimates $\hat{\theta}$ and normalised by $\hat{r}$. Hence values larger than one are not small enough. The second summand is proportional to the squared Euclidean norm of $({}^c\psi_t - {}^c\,\psi_\kappa)$ weighted by the precision matrix C^{-1}, which can be described as an inversion of the second sample moment of the regression vectors divided by ν. Hence, the values of divergence

$$\mathcal{D}_{t\kappa} \leq \frac{1}{2}(1 + l_\psi/\nu) \quad (17)$$

can be considered sufficiently small, [3].

Thus to learn a model of the environment at time t, we search for the relevant regression vectors ψ_κ via (16) and use them to build model at time t according to LS algorithm described in Section 2.2. This model is then used as model of the environment.

4 Results

General aim of this part is to illustrate the proposed approach combining LL and structure and parameter estimation. The task is to learn the dependencies in the whole system and to make high-quality prediction of the environment behaviour, which can be further used in the optimisation. The behaviour of the environment changes with time and may have different modes. So it can be described by a collection of several models, each valid at some time period. We learn the main dependencies between data and estimate the relevant coefficients of this structure. Eventually, we explore the ability to predict the future behaviour of the environment.

The experimental data are taken from International Fast Moving Consumer Goods company[1]. The data containing 30 variables has been observed for 139 weeks. The observed environment states characterise sale of selected goods. Other variables potentially entering regression vector contain price, sales volumes, marketing indicators (leaflets, TV commercials, etc.) and index of the location of the goods in the store.

[1]Any details about data and company are subject to anonymity restrictions.

4.1 Algorithm

We use the logarithm of sales values as s_t and work with log-normal model. The maximal delay N (see Definition 1) is set to one week. It is assumed that older data do not influence the present behaviour. The overall algorithm implementing the proposed approach is as follows.

1. At the very beginning Bayesian *approach without lazy learning* is used to analyse the data set. At each time step: (i) structure of the model (2) is estimated; (ii) corresponding estimates of regression coefficients, $\hat{\theta}$, (5) are updated and, (iii) prediction of the environment state (15) is made. Recall that the model is log-normal, the mean is $\exp(\mu + \frac{r}{2})$, where μ is the mean and r is the variance of the related normal distribution $\mathcal{N}(\mu, r)$. Prediction is then equal $\hat{s}_t = \hat{\theta}_{t-1}\psi_t$.

2. Next step considers separating the past data into several sets, each corresponds to one working mode (described by an individual model). Here due to the lack of data we consider only two working modes. To make the separation, every past regression vector ψ_κ, $\kappa = \tau, \tau+1, \ldots, t-1$, where τ is learning time, is compared with the current regression vector, ψ_t via corresponding KLD. Having the learning time τ fixed, the computed values of KLD, (16), are $\mathcal{D}_{t\kappa}, \kappa = \tau, \tau+1, \ldots, t-1$. By setting a boundary value $d > 0$, we can separate the original data set into two sets: *similar data*, i.e. data with regression vectors close to the actual one, ψ_t, $\mathcal{D}_{t\kappa} < d$, and data with regression vectors having KLD $\mathcal{D}_{t\kappa} \geq d$.

   A boundary value d is defined such that if $\mathcal{D}_{t\kappa} < d$, the regression vector at time κ belongs to one working mode, otherwise ($\mathcal{D}_{t\kappa} \geq d$) it belongs to another. Hence, at each time step only one of the modes (described either by model $M1$ or $M2$) is valid. Boundary value of KLD d is set empirically, based on (17).

3. The structure of each model ($M1$ and $M2$) is estimated using the available data.

4. Finally, we make predictions based on the *approach with lazy learning*. At each time we compare actual data with all past data, i.e. calculated the KLD, $\mathcal{D}_{t\kappa}(\psi_t||\psi_\kappa)$, and searched for the regression vector closest to the current one. The prediction is made based on the relevant structure at some past time τ, where $\tau = \text{argmin}_\kappa(\mathcal{D}_{t\kappa})$.

After the predictions with LL and without LL are obtained, we calculate the prediction errors $e_\tau = s_\tau - \hat{s}_\tau$, where s_τ is the real state and $\hat{s}_\tau$ is the prediction of this state. To compare the approaches, the error is calculated as root-mean-square error of the prediction normalized by the root-mean measured value.

These two approaches are confronted with *trivial prediction*. Trivial prediction means that the environment state at time $t+1$ is going to be the same as already observed state at time t, i.e. $\hat{s}_{t+1} = s_t$.

4.2 Examples

Throughout the experiments the special notations connected with software implementation are used. *Channel* indicates the position of particular variable in the maximal

regression vector. *Delay* can be 0 or 1 week, because the maximal delay (N) of variables present in the regression vector is set to 1 week, see (2). That means at time t, $s_{9,t}$ stands for the sales stored in Channel 9 with zero delay and $s_{9,t-1}$ for one-week delay. The initialisation (learning time) for the predictions is set to 30 as we have small amount of data. Recall that the data contain 30 variables, i.e. price, sales volumes, marketing indicators (leaflets, TV commercials, etc.) and index of the location of the goods in the store.

Consider selling Product 1 (P1) by a Seller 1 (S1). Our aim is to predict sales values. First we applied *approach without lazy learning*, i. e. we learned the model structure and regression coefficients based on the whole data set. The results are in Table 1.

Channel	16	9	21	8	26	16
Delay	1	1	0	1	0	0
Coefficient	-1.0524	8.8417	-1.5553	-6.931	-1.4128	2.5996

Table 1: Model structure, corresponding regression coefficients and delays (approach without LL).

The corresponding model is then

$$s_{9,t} = -6.93s_{8,t-1} + 8.84s_{9,t-1} + 2.60a_{16,t} - 1.05a_{16,t-1} - 1.56a_{21,t} - 1.41a_{26,t} + \hat{e}_t,$$

where $s_{9,t}$ stands for the sales values and $\hat{e}_t$ is the prediction error in the actual time t, see Section 4.1.

Then we divided the whole data set into two and learned respective models $M1$ and $M2$. Table 2 shows the model structures and corresponding estimates of regression coefficients. Then *approach with lazy learning* is used. At each time t we compare the actual regression vector with all past vectors (see Section 3.2). The prediction is made based on the model ($M1$ or $M2$) corresponding to the closest regression vector ψ_i, where $i = \text{argmin}_\kappa(\mathcal{D}_{t\kappa})$, see Section 4.2.

	Model M1			Model M2			
Channel	26	16	21	7	21	27	16
Delay	1	0	0	0	0	1	0
Coefficient	-0.8884	2.6415	-1.9270	0.8554	-1.9990	-1.1612	2.4277

Table 2: Model structure, corresponding regression coefficients and delays (approach with LL).

In a similar way we proceeded two other examples describing selling other products. Table 3 compares the prediction error of trivial prediction, approach without lazy learning and our proposed approach for all three examples. The improvement brought by LL is significant compare to other approaches.

4.3 Discussion

An assumption, that data older that 1 week (maximal delay) do not influence the actual sales, was confirmed experimentally.

Due to the lack of data, the separating value d set according to (17), where $l_\psi = 30, \nu = 32$, $d \approx 1$, was not good.

The separating value is set based on two demands. Firstly, the subsystems have to be valid for longer time than the learning time (30). Then, we focus on the visual criterion. As the businesses data are proceeded, we can recognize different subsystems, selling strategies, data collection mistakes etc. The division based on KLD has to match with the reality. Empirically selected value $d = 2$ provided the prediction of good quality. The reason for that might be the following. Both time curve of KLD values and time curve of sales values have big peaks, which may represent external influences, for instance promotions and discounts, which may cause significant increase of sales. Approaches with and without LL provide similar quality predictions when sales curves do not have peaks, but LL approach significantly better predicts peaks in sales.

Let us compare Table 1 and Table 2. Table 1 describes model learned on the all data. Then original data were divided into two sets and Table 2 describes two models, each learned on respective part of the original data set. For instance, compare to the model learned on the whole data (Table 1), Model $M2$ (Table 2) contains two new variables (Channel 7 and Channel 27). Channel 7 reflects whether leaflet was used or not. As leaflets usually announce promotions, model $M2$ may describe the sales behaviour during promotion. This strongly supports our motivation to separate the original data set.

We add the comparison of the approaches for Example 2 and Example 3, describing sales of P2 and P3 by S2 and S3 (Table 3). There were many promotions during the 139 weeks in the Example 2. Almost every promotion or advertising is represented by a peak in sales. Our approach improved the predictions of these peaks. However, the improvements are smaller than in Example 1. Recall that during the basic sale the predictions differ only a little and that is why we focus on the peaks as the most interesting part for the seller performance. The Example 3 shows that separating into two subsystems may not be sufficient. We did many improvements and degradation of the prediction during the time and that could be solved by more than two separating levels. We tried to set two separating values d_1, d_2 to describe by three models. However, the prediction was worse due to low tolerance in changing of d_1, d_2 and there was no way to improve this due to the lack of data.

	Example 1	Example 2	Example 3
Error of trivial prediction	0.9664	0.9132	0.4647
Error of prediction without LL	0.6061	0.6512	0.3278
Error of prediction with LL	0.4080	0.5848	0.3421

Table 3: Errors of the different approaches.

5 Conclusion

The LL approach for learning environment model is proposed and tested on real marketing data. The key idea is to describe the overall process by several models, each describing one operating mode. The separation of the data depends on selecting a separating value d. Three different examples were considered, Chapter 4.1. The main difficulty is small

amount of data available. Examples considered have significant differences in the structures, which calls for large sets of data. The disadvantage of the proposed approach is possible reaction on the environment dynamics and transitions between the different modes. Errors occur during the transitions, if the states change too fast. Hence, the suggested form of lazy learning is a powerful tool for dealing with long term errors in data collection. The approach able to find and emphasize new significant variables, that were not included in the single-model structure.

The proposed approach significantly improves the prediction quality in the most experiments. The choice of the value d is important and determines how many modes (models describing the environment) should be considered. As mentioned, the separation is based on the value of KLD, see Section 3.2. The approach is very sensitive to the choice of d. A slight change of d causes big changes in number and structure of models. The algorithm how to properly select value of d remains one of the most demanding open questions. Another open question is a proximity measure. Here we use KLD but there is no evidence that this is the best practical choice.

Acknowledgement: This research has been partially supported by GAČR 13-13502S and Contract 104500-AS.

References

[1] M. Kárny, J. Bohm, T. V. Guy, L. Jirsa, I. Nagy, P. Nedoma and L. Tesař *Optimized Bayesian Dynamic Advising: Theory and Algorithms*. Springer, (London 2006).

[2] M. Kárny, J. Andrýsek, A. Bodini, T. V. Guy, J. Kracík, F. Ruggeri : How to exploit external model of data for parameter estimation? *International Journal of Adaptive Control and Signal Processing vol.20*, p. 41-50, 1 (2006).

[3] V. Peterka *Bayesian system identification.*. Eykhoff P. (Ed.), Trends and Progress in System Identification. Pergamon Press, Oxford, pp. 239-304. 1981, MR0746139

[4] C. G. Atkeson, A. W. Moore, and S. Schaal. *Locally weighted learning for control*. Artif. Intell. Rev., 11(1-5):75-113,1997.

[5] K. Macek, T. V. Guy, M. Kárný *A Lazy-Learning Concept of Fully Probabilistic Decision Making*. Artificial Intelligence, under revision [2014]

[6] M. Kárný. *Algorithms for determining the model structure of a controlled system*. Kybernetika, 19(2): 164-178, 1983.