EasyChair Preprint

№ 8303

Continuous Model Evaluation and Adaptation to Distribution Shifts: a Probabilistic Self-Supervised Approach

Gregor Pavlin, Pieter de Villiers, Kathryn Laskey, Franck Mignet and Lennard Jansen

June 18, 2022

Continuous Model Evaluation and Adaptation to Distribution Shifts: A Probabilistic Self-Supervised Approach

G. Pavlin[‖] J. P. de Villiers[*] K. Laskey[‡]
F. Mignet[‖] L. Jansen[‖]
[‖]Thales Research and Technology, Delft, The Netherlands, gregor.pavlin@nl.thalesgroup.com
[*] University of Pretoria, Pretoria, South Africa, pieter.devilliers@up.ac.za
[‡]George Mason University, Fairfax, VA, USA, klaskey@gmu.edu

***Abstract*—This paper introduces a Bayesian approach to estimating distribution shifts over the modelled variables and continuous model adaptations to mitigate the impact of such shifts. The method exploits probabilistic inference over sets of correlated variables in causal models describing data generating processes. By extending the models with latent auxiliary variables, probabilistic inference over sets of correlated variables enables estimation of the distribution shifts impacting different parts of the models. Moreover, the introduction of latent auxiliary variables makes inference more robust against distribution shifts and supports automated, self-supervised adaptation of the modelling parameters during the operation, often significantly reducing the adverse impact of the distribution shifts. The effectiveness of the method has been validated in systematic experiments using synthetic data.**

***Index Terms*—distribution shift, inference, Bayesian Networks, machine learning, trust**

I. Introduction

Advances in machine learning have enabled applications that rely on automated classification and prediction using complex data patterns. The estimation of the expected accuracy of such solutions under operational conditions is a key in building trust. In machine learning it is typically assumed that the data used for the training of the models and the data used at runtime are sampled from the same probability distribution $P(\mathcal{V})$, where $V = \{v_1, v_2, \ldots, v_N\}$ denotes the set of relevant random variables. However, this is often not a realistic assumption. Typically, the training data is sampled from $P(\mathcal{V})$ while the data used at runtime is sampled from a different distribution $P(\mathcal{V})^*$, i.e. Kullback-Leibler divergence $D_{KL}(P(V)^*||P(V)) \neq 0$. In such a case, the actual accuracy of the classifiers and predictors can be significantly lower than the expected accuracy obtained during the testing of the models using the data sampled from $P(\mathcal{V})$. This paper introduces a probabilistic approach that supports identification of distribution shifts at runtime and can use this information to automatically correct the model reducing the adverse effects of such shifts. The method is based on a causal modelling pattern that enables probabilistic inference about distribution shifts with respect to different components of a probabilistic model. The approach described in this paper partially builds on the ideas presented in [1], where the authors utilised latent variables to model changes in the quality and relevance of sensor data. That approach, however, focused on the analysis of data sources and could handle models consisting of binary variables only.

The approach presented in this paper, on the other hand, exploits correlations between different observations to **pinpoint any modelling components** that have been rendered sub optimal due to the changes in the corresponding probability distributions characterizing the true data generating process. The presented approach introduces multiple benefits. Firstly, it enables identification of inadequate components in factorized probabilistic models of data generating processes. If the distribution shifts in the true data generating processes are limited to small subsets of variables and occur over extended periods of time, the identified modelling components could be repaired with data sets that are small compared to the sets required for relearning of the entire model. Secondly, the introduced modelling pattern improves inherent robustness against the distribution shifts as it reduces their impact on the overall inference process, i.e. it acts as a "shock absorber" in the reasoning process. Thirdly, the pattern supports implementation of auto correcting mechanisms that exploit estimation of biases and use this information to partly compensate the impact of distribution shifts.

A. Related work

The concept of data shift, or dataset shift, has seen increased interest since approximately 2008 [2]–[4]. Some of the early references to the dataset shift problem can be found in [5]. A standard definition of dataset shift [6] is:

Definition I.1 (Dataset shift)**.** Dataset shift appears when training and test set probability distribution functions are different. That is when $P_{tr}(x, y) \neq P_{tst}(x, y)$, where x denotes the set of covariates or features and y denotes the target variable.

Before a standard definition of dataset shift was presented, other terminology was used to describe the same phenomenon. Alternate terms include concept shift, changes of classification,

changing environments, contrast mining, fracture points and data fractures [6]. Dataset shift can also take different forms, namely

1) *covariate shift*, where $P_{tr}(y|x) = P_{tst}(y|x)$ and $P_{tr}(x) \neq P_{tst}(x)$ (appearing in problems where the class label is causally determined by the covariate values – so called $X \rightarrow Y$ problems),
2) *prior shift*, where $P_{tr}(x|y) = P_{tst}(x|y)$ and $P_{tr}(y) \neq P_{tst}(y)$ (appearing in problems where the covariate values are causally determined by the class label – so called $Y \rightarrow X$ problems),
3) *concept shift*, which can be summarized as changes to the definition of the classes, i.e. $P_{tr}(y|x) \neq P_{tst}(y|x)$, $P_{tr}(x) = P_{tst}(x)$ in $X \rightarrow Y$ problems, and $P_{tr}(x|y) \neq P_{tst}(x|y)$, $P_{tr}(y) = P_{tst}(y)$ in $Y \rightarrow X$ problems,
4) *other types* not occurring regularly in most practical problems, where $P_{tr}(y|x) \neq P_{tst}(y|x)$, $P_{tr}(x) \neq P_{tst}(x)$ in $X \rightarrow Y$ problems, and $P_{tr}(x|y) \neq P_{tst}(x|y)$, $P_{tr}(y) \neq P_{tst}(y)$ in $Y \rightarrow X$ problems.

Of particular interest is non-stationary environments, where the joint class-data distributions change with space or time. The obvious application where this takes place is adversarial classification (for example fraud detection, spam filtering and intrusion detection [6]). Other examples include condition-based monitoring in mechanical systems, where parts "wear out" over time [7], [8] and in the case of sensor measurements, changes in the environmental conditions [4].

The first investigation into dataset or distribution shifts within the context of data fusion can be found in [9] within an application of identifying failed sensors. Shifts in the distribution of the data have important implications when considering the representation and reasoning in a fusion system. Over several years, the Evaluation Techniques for Uncertainty Reasoning Working Group (ETURWG) has been developing the Uncertainty Representation and Reasoning Framework (URREF) ontology that can be used as a framework to evaluate uncertainty representation and reasoning in a fusion system [10]–[14]. Dataset shift influences both the uncertainties in representation (model) of the fusion system as well as the reasoning (inference algorithm) aspects of the fusion system. Representation uncertainties are compounded when the system is designed and trained for an environment which is different from the environment in which the system is deployed. This contributes to challenges of certifiability of the deployed system, in that certain performance parameters cannot be guaranteed where some environmental, sensor or system parameters fall outside those for which the system was trained or tested [15]–[18]. Reasoning uncertainties are compounded by dataset shift through, for example in a Bayesian context, a mismatch of support between the prior distribution and the likelihood function, leading to more diffuse posterior distributions (see section II-A).

The consequences of dataset drift can be as a result of the conceptualisation, design, implementation, testing and operational phases of fusion system development [19]. Dataset distribution shifts can already manifest during the conceptualisation phase owing to misguided characterisation of fusion inputs/outputs, and/or selection of a mismatched representation and reasoning schemes. Training typically takes place during the design phase of the fusion system. Dataset shift can be introduced during training already, by not training on representative data, or training on synthetic data, in cases where measured data is scarce. During fusion runtime, dataset distribution shifts are encountered, owing to the system encountering data which has changed, or situations with outcomes that are slightly different to those for which the system was designed.

The relevant URREF criteria that relate to uncertainties owing to dataset shift related to representation (modelling) are subclasses under the *RepresentationCriterion*, where the subclasses of *Expressiveness*, *Simplicity*, *Adaptability* and *Compatibility* are relevant uncertainty evaluation cirteria. More notably, given the unknown (epistemic) aspects of dataset shift owing to a variety of conceptual variations in real world applications, evaluation of *HigherOrderUncertainty*, i.e. uncertainty about uncertainty is a particularly relevant evaluation criterion.

The URREF criteria that relate to uncertainties owing to dataset shift related to reasoning (inference algorithms) are subclasses under the *ReasoningCriterion*. Here, *Consistency*, *Correctness*, *Scalability* and *Performance* are relevant criteria. It is clear that all of the mentioned criteria will be affected by dataset shift. Apart from the representation and reasoning criteria, there are obviously subclasses of the *DataCriterion* which are criteria relevant to dataset shift [14]. These include *DataQuality* criteria which include *Accuracy* and *Precision*, as *RelevanceToProblem* and *WeightOfInformation* (a.k.a. *WeightOfEvidence* in some versions of the ontology).

The paper is organized as follows: Section II introduces the modelling pattern and its use; section III discusses how to detect dataset shift and estimate its magnitude; section IV describes the self-correcting inference processes; and section V presents the experiments and discusses the results. Finally, section VI describes applications of the method, and section VII provides concluding remarks.

II. Modelling

We assume a causal probabilistic model that encodes the joint probability distribution $P(\mathcal{V})$ over a set of random variables $\mathcal{V} = \{v_1, v_2, \ldots, v_N\}$. The model represents a data generation process producing observations ϵ produced by different sources (e.g. sensors) that are correlated with the states of the variables of interest. Moreover, we are assuming that the model captures the joint probability/density over all relevant variables $\mathcal{V}$ that can be represented through a factorization:

$$P(\mathcal{V}) = \prod_{v_i \in \mathcal{V}} P(v_i|\pi(v_i)), \tag{1}$$

where $\pi(v_i)$ denotes the set of all parents of v_i, i.e. the variables that directly influence the states of v_i. For root variables $\pi(v_i) = \emptyset$ and $P(v_i|\emptyset) = P(v_i)$. Such factorization corresponds to a graph representing direct dependencies between variables in $\mathcal{V}$. An example of such a graph is

shown in Fig. 1. The graph corresponds to the factorization of the joint probability distribution over the set of variables $\mathcal{V} = \{Class, O, R, E, A, B, C\}$:

$$P(\mathcal{V}) = P(Class)P(O|Class)P(R|Class) \\ P(E|Class)P(A|E)P(B|E)P(C|E). \quad (2)$$

In this paper it is assumed that the parameters in (2) are obtained through machine learning, such as Expectation Maximization (EM), as some variables are latent, i.e. their states are not observed during the sampling of the training data. E in (2) is an example of a latent variable. This is an example of an $Y \rightarrow X$ problem, where the "features" are conditioned on the class variable. Dataset shift can occur in two ways. The first is through changes to the prior class distribution through $P_{trn}(Class) \neq P_{tst}(Class)$, i.e. prior shift. Alternatively, there may be changes to the conditional distributions of the child variables given the class, for example $P_{trn}(E|Class) \neq P_{tst}(E|Class)$, i.e. concept shift.

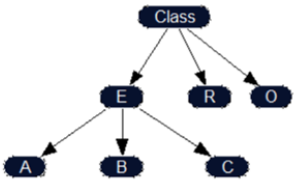

Figure 1: A basic model describing correlations between the variables in a data generation process.

A. Impact of Distribution Shifts

The factorization in (1) is key to efficient inference over the states of any unobserved variable $v_i \in \mathcal{V}$ given a set of observations ϵ, where the symbol ϵ represents evidence, i.e., assignment of definite values to a subset of the variables in $\mathcal{V}$. Probabilistic inference is the computation of $P(v_i|\epsilon)$, the posterior distribution over the states of an unobserved variable v_i using Bayes rule[1]

$$P(v_i|\epsilon) = \\ \eta \int_{\pi(v_i)} P(\pi(v_i)|\epsilon_{pa})P(v_i|\pi(v_i))P(\epsilon_{ch}|v_i)d\pi(v_i), \quad (3)$$

where η is a normalizing constant, ϵ_{pa} denotes the set of observations that can be reached from v_i by traversing the graph via its $\pi(v_i)$ and ϵ_{ch} denotes the set of observations that can be reached by traversing the graph via v_i's children $ch(v_i)$; i.e. the total set of observations[2] $\epsilon = \epsilon_{pa} \cup \epsilon_{ch}$. Equation (3) fuses evidence about causes of v_i (variables upstream in the graph) with evidence about effects of v_i (variables downstream in the graph) via the conditional probability or density $P(v_i|\pi(v_i))$. If $P(v_i|\pi(v_i))$ correctly represents the true distribution, then $P(v_i|\epsilon_{pa}) = \int_{\pi(v_i)} P(v_i|\pi(v_i))P(\pi(v_i)|\epsilon_{pa})d\pi(v_i)$ will support the same states of v_i as its likelihood $P(\epsilon_{ch}|v_i)$ based on the observations ϵ_{ch}. In other words, $\arg\max_{v_i} P(v_i|\epsilon_{pa})$ is close to $\arg\max_{v_i} P(\epsilon_{ch}|v_i)$. Such a situation is illustrated in Fig. 2.a).

If the true conditional distribution of v_i given its parents $\pi(v_i)$ changes, i.e. $P(v_i|\pi(v_i)) \neq P(v_i|\pi(v_i))^*$, then $\arg\max_{v_i} P(v_i|\epsilon_{pa})$ may be significantly different from $\arg\max_{v_i} P(\epsilon_{ch}|v_i)$, resulting in low $P(v_i|\epsilon)$. Such a situation is illustrated in Fig. 2.b). Namely, given $P(\pi(v_i)|\epsilon_{pa})$, the distribution $P(v_i|\pi(v_i))$ suggests a different sample ϵ_{ch} [3].

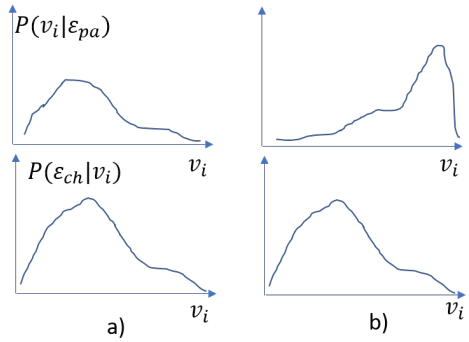

Figure 2: a.) $P(v_i|\epsilon_{pa})$ and $P(\epsilon_{ch}|v_i)$ support similar states of v_i. b.) $P(v_i|\epsilon_{pa})$ and $P(\epsilon_{ch}|v_i)$ support different states of v_i.

B. Extended Models

To represent a shifted $P(\mathcal{V})^*$, a new model is obtained by extending the original model. This is achieved by (i) adding a set of auxiliary variables $\mathcal{D} = \{\Delta v_1, \Delta v_2, \ldots, \Delta v_M\}$, resulting in a new set of model variables $\mathcal{V}' = \mathcal{D} \cup \mathcal{V}$ and (ii) replacing the conditional distribution $P(v_i|\pi(v_i))$ of each variable v_i by $P(v_i|\pi(v_i), \Delta\pi(v_i))$, where $\Delta\pi(v_i) \subseteq D$ denotes the set of auxiliary variables that are additional parents of variable v_i. This leads to a new factorization:

$$P(\mathcal{V}') = \prod_{\Delta v_i \in \mathcal{D}} P(\Delta v_i) \prod_{v_i \in \mathcal{V}} P(v_i|\pi(v_i), \Delta\pi(v_i)). \quad (4)$$

This factorization corresponds to a modified graph. Fig. 3 shows an example model that was obtained in this way from the initial model in Fig. 1. Variables with prefix Δ denote auxiliary variables and form the auxiliary set $\mathcal{D}$. A distribution shift between $P(v_i|\pi(v_i))$ and $P(v_i|\pi(v_i))^*$, can be seen as a function of the states of v_i's auxiliary variables $\Delta\pi(v_i)$. By controlling the states of $\Delta\pi(v_i)$, $P(v_i|\pi(v_i))^*$ is obtained, i.e. $P(v_i|\pi(v_i), \Delta\pi(v_i)) = P(v_i|\pi(v_i)^*$.

$P(v_i|\pi(v_i), \Delta\pi(v_i))$ could take on different forms. Parametric distributions are especially useful, as they simplify modelling of the distribution shifts. For example, let's assume $P(v_i|\pi(v_i))$ is a Gaussian distribution $P(v_i|\pi(v_i)) \sim \mathcal{N}(\mu_{\pi(v_i)}, \sigma^2_{\pi(v_i)})$, where $\mu_{\pi(v_i)}$ and $\sigma_{\pi(v_i)}$ are parameters

[1]For the sake of brevity all equations in this paper assume continuous variables and probability densities. The same equations are valid for discrete variables, in which case the integrals are replaced with sums.

[2]Equation (3) assumes that v_i d-separates ϵ_{pa} from ϵ_{ch}, which is always the case if the graph is singly connected.

[3]The observations ϵ_{ch} can be viewed as a sample from $P(\epsilon_{ch}|v_i)$. If $P(v_i|\pi(v_i)) \neq P(v_i|\pi(v_i))^*$, then ϵ_{ch} is conditioned on different states of v_i than assumed according to $P(v_i|\epsilon_{pa})$ based on the original $P(v_i|\pi(v_i))$.

that depend on the states of v_i's parents $\pi(v_i)$ and were learned from the data sampled on the original distribution $P(\mathcal{V})$. By adding auxiliary variables $\Delta pa(v_i)$ we can obtain $P(v_i|\pi(v_i), \Delta\pi(v_i))$ as follows

$$P(v_i|\pi(v_i), \Delta\pi(v_i)) \sim \mathcal{N}(\mu_{\pi(v_i)} + \Delta\mu_{\Delta\pi(v_i)}, \sigma^2_{\pi(v_i)}),$$

where $\Delta\mu_{\Delta\pi(v_i)}$ represents a distribution shift as a function of the states of the v_i's auxiliary variables. This equation can be easily extended also by making $\sigma_{\pi(v_i)}$ a function of $\Delta\pi(v_i)$. Other types of more advanced parametric distributions can be used to represent more elaborate distributions for conditional probabilities, such as Gaussian and other types of mixtures.

In the example model shown in Fig. 3 it is assumed that E is a latent continuous random variable whose $P(E|Class, \Delta\pi(E))$ is defined through three Gaussian components, each associated with one of the states of the discrete $Class$ variable. Continuous variables $\Delta\pi(E) = \{\Delta Ea, \Delta Eb, \Delta Ec\}$ denote the shifts of the means for these components, such that: $P(E|Class = a, \Delta\pi(E)) \sim \mathcal{N}(\mu_a + \Delta Ea, \sigma^2_a)$, $P(E|Class = b, \Delta\pi(E)) \sim \mathcal{N}(\mu_b + \Delta Eb, \sigma^2_b)$ and $P(E|Class = c, \Delta\pi(E)) \sim \mathcal{N}(\mu_c + \Delta Ec, \sigma^2_c)$. Continuous variables A, B and C, on the other hand, represent sensor measurements whose models are defined over E's values as follows:
$P(A|E, \Delta\pi(A)) \sim \mathcal{N}(\mu_E + \Delta A, \Delta As^2_E)$, $P(B|E, \Delta\pi(B)) \sim \mathcal{N}(\mu_E + \Delta B, \Delta Bs^2_E)$ and $P(C|E, \Delta\pi(C)) \sim \mathcal{N}(\mu_E + \Delta C, \Delta Cs^2_E)$. Note that each of these variables has two auxiliary variables defining the shift and the standard deviation respectively, i.e. $\Delta\pi(A) = \{\Delta A, \Delta As\}$, $\Delta\pi(B) = \{\Delta B, \Delta Bs\}$ and $\Delta\pi(C) = \{\Delta C, \Delta Cs\}$. Moreover, the discrete $Class$ variable has three states and is also influenced by continuous auxiliary variables $\Delta\pi(E) = \{\Delta Class_a, \Delta Class_b, \Delta Class_c\}$ representing parameters of a discrete categorical distribution. Discrete variables O and R, on the other hand, represent outputs of two different types of detectors that can classify the states of $Class$. These two variables are assumed to be influenced also by auxiliary variables $\Delta\pi(O) = \{\Delta O1, \Delta O2, \Delta O3\}$, and $\Delta\pi(R) = \{\Delta R1, \Delta R2, \Delta R3\}$, respectively. $\Delta\pi(O)$ and $\Delta\pi(R)$ denote shifts of the categorical distribution over three types of classifications used in the conditional probability tables of the two discrete variables.

III. Estimating Distribution Shifts

The aforementioned model extensions enable estimation of the distribution shifts for each variable v_i, i.e. deviations between the originally learned parameters $P(v_i|\pi(v_i))$ and the current distribution $P(v_i|\pi(v_i))^*$ influencing the data generating processes.

$P(v_i|\pi(v_i), \Delta\pi(v_i))$ defines a set of different distributions corresponding to possible situations determined through the states of $\pi(v_i)$ and $\Delta\pi(v_i)$. For the extended model, the distribution $P(v_i|\epsilon)$, is computed using Bayes rule and integrating over the auxiliary variables as follows:

$$P(v_i|\epsilon) =$$

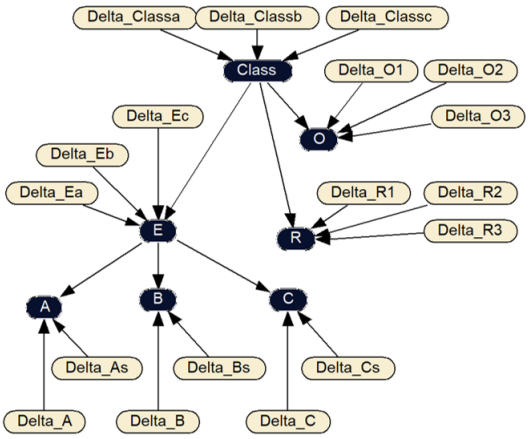

Figure 3: An extended model supporting inference about the distribution shifts. Yellow nodes represent auxiliary variables.

$$\eta_v \int_{\pi(v_i)} \int_{\Delta\pi(v_i)} [P(\Delta\pi(v_i))P(\pi(v_i)|\epsilon_{pa}) \times P(v_i|\pi(v_i), \Delta\pi(v_i))P(\epsilon_{ch}|v_i)] d\pi(v_i) d\Delta\pi(v_i). \quad (5)$$

The magnitude of the distribution shift can be estimated using inference. For any auxiliary variable $\Delta v_j \in \Delta\pi(v_i)$ Bayes rule can be used to compute the distributions over its states, given all observations ϵ:

$$P(\Delta v_j|\epsilon) = \eta_\Delta \int_{v_i} \int_{\pi(v_i)} \int_{\Delta\pi(v_i)\setminus\Delta v_j} [P(\Delta\pi(v_i))P(\pi(v_i)|\epsilon_{pa}) \times P(v_i|\pi(v_i), \Delta\pi(v_i))P(\epsilon_{ch}|v_i)] dv_i d\pi(v_i) d\Delta\pi(v_i). \quad (6)$$

This operation estimates the distribution over the states of Δv_j based on the discrepancy between $P(\pi(v_i)|\epsilon_{pa})$ and $P(\epsilon_{ch}|v_i)$ for different combinations of states of $\pi(v_i)$ and $\Delta\pi(v_i) \setminus \Delta v_j$. $P(\Delta v_j|\epsilon)$ increases for non-zero shifts if $\arg\max_{v_i} P(\epsilon_{ch}|v_i)$ is significantly different from $\arg\max_{v_i} P(v_i|\epsilon_{pa})$.

Equation (6) enables estimation of the shift of different components of the model. $P(\Delta v_i|\epsilon)$ can be used as an indicator for the occurrence of distribution shifts, alerting the user about potentially degraded system performance.

IV. Robustness and Self-Supervised Model Adaptation

The extended models presented in Section II-B enable robust inference in the presence of distribution shifts. By incorporating the possibility of distribution shifts within a Bayesian framework, the estimate $P(v_i|\epsilon)$ given in Equation (5) is inherently robust to distribution shifts. Reasoning based on $P(v_i|\pi(v_i), \Delta\pi(v_i))$ allows an alternative explanation for significant differences between $\arg\max_{v_i} P(\epsilon_{ch}|v_i)$ and $\arg\max_{v_i} P(v_i|\epsilon_{pa})$. In case of such differences, the reasoning increases the probability of the states of Δv_i that would explain the discrepancy. **Thus, the reasoning with auxiliary Δv_i variables reduces the impact of inadequate**

modelling parameters; these variables have a role of "shock absorbers" in the inference process.

Auxiliary variables enable automatic correction of the models at runtime. This is achieved by creating composite data generation models representing N events in the same domain in which sets of observations $\epsilon_{1:N} = \epsilon_1, \ldots, \epsilon_N$ were obtained. The extended model whose graph is shown in Fig. 3 is an example of the description of the data generation process in a single event. For N estimation events, all variables from such an extended model are replicated N times, except the auxiliary variables in set $\mathcal{D}$, as it is assumed that distribution shifts had not changed during the period in which N estimation events took place. Consequently, a single set $\mathcal{D}$ is used throughout all events. The topology of an example composite model is shown in Fig. 4.

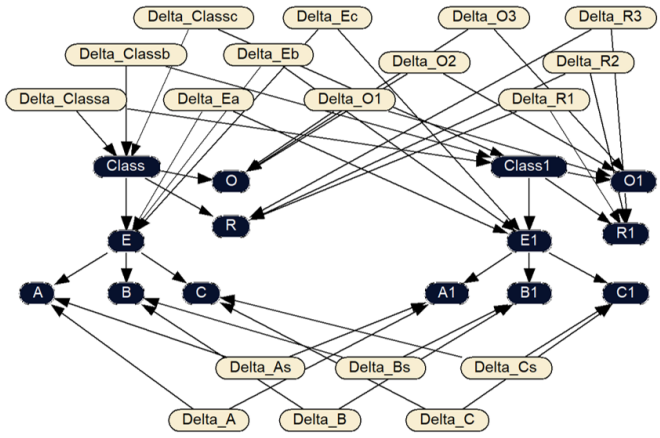

Figure 4: A model supporting inference about the distribution shifts over two estimation events. Auxiliary shift variables are represented by yellow nodes.

Such composite models can support effective mitigation of the impact unknown distribution shifts have on inference processes. For example, in a class estimation process at the N^{th} estimation event using observations ϵ_N, the computation of class posteriors can be improved by using observations $\epsilon_{1:N-1}$ from the preceding $N-1$ events to estimate the states of auxiliary variables in $\mathcal{D}$. Conditional probabilities $P(Class_N|\epsilon_{1:N})$ and $P(\Delta v_j|\epsilon_{1:N})$ for the N^{th} can be computed by using well known algorithms supporting Bayesian inference in multi-loop Bayesian networks, such as the Junction Tree [20]. Without discussing the details of this sophisticated algorithm, the benefits from using the multi-event models can be shown by transforming the original network through grouping of variables, such that all auxiliary variables in $\mathcal{D}$ are considered a single variable D. $P(Class_N|\epsilon_{1:N})$ can then be expressed as

$$P(Class_N|\epsilon_{1:N}) = \eta_N \int_D P(Class_N|D)P(\epsilon_N|Class_N, D)P(D|\epsilon_{1:N-1})dD \tag{7}$$

where $P(D|\epsilon_{1:N-1})$ denotes the estimated joint distribution over all auxiliary variables in set $\mathcal{D}$ conditioned on the observations from $N-1$ preceding events. As the model was obtained through replication, $P(Class_N) = P(Class_{N-1} = \ldots = P(Class_1))$ and $P(\epsilon_1|Class_1, D), \ldots, P(\epsilon_N|Class_N, D)$ are based on identical conditional probability tables. By inspecting the graph in Fig. 4, we see that auxiliary variables in $\mathcal{D}$ form a d-separation set [20] between the replicated extended models. Therefore $P(D|\epsilon_1, \ldots, \epsilon_{N-1})$ can be expressed as a factorization

$$\begin{aligned} P(D|\epsilon_{1:N-1}) = \eta_D P(D)[\\ \sum_{Class_1} P(Class_1|D)P(\epsilon_1|Class_2, D)\times \\ \sum_{Class_2} P(Class_2|D)P(\epsilon_2|Class_2, D)\times \\ \ldots \\ \sum_{Class_{N-1}} P(Class_{N-1}|D)P(\epsilon_{N-1}|Class_{N-1}, D)], \end{aligned} \tag{8}$$

where $P(\epsilon_i|Class_k, D)$ denotes a mapping that is obtained from the k-th replicated part of the model. As $\mathcal{D}$ d-separates the modelling fragments corresponding to different estimation events, this computation can be carried out in a recursive manner. The auxiliary variables can be seen as memory elements, capturing estimates of the shifts from preceding estimation events. The observations from different events are correlated via the auxiliary variables in $\mathcal{D}$. Therefore, the classification process in the N^{th} step is supported not only by the current observations ϵ_N but also by the preceding sets of observations $\epsilon_{1:N-1}$. This is due to the inference processes estimating $P(D|\epsilon_{1:N-1})$. As more events are considered, the estimate $P(D|\epsilon_{1:N-1})$ is likely to approach the true distribution shifts, improving the estimation $P(C_N|\epsilon_{1:N})$ in the process.

Moreover, $P(\Delta v_j|\epsilon_{1:N})$ can be obtained through marginalization of $P(D|\epsilon_{1:N})$

$$P(\Delta v_j|\epsilon_{1:N}) = \eta_D \int_{\mathcal{D}\setminus\Delta v_j} P(D|\epsilon_{1:N})dD \tag{9}$$

V. Experiments

Multiple experiments were carried out to validate the approach using extended models with different numbers of estimation events. Fig. 4 shows an example with two estimation events. The synthetic data for experiments was prepared by (i) using an extended model representing 30 estimation events, (ii) splitting auxiliary variables between sets $\mathcal{D}_s$ and $\mathcal{D}_0$, such that $\mathcal{D}_s \cup \mathcal{D}_0 = D$ and (iii) fixing the values of variables in $\mathcal{D}_0$ at zero shift. For each combination of $\mathcal{D}_s$ and $\mathcal{D}_0$, the model was used to produce 3000 samples, each consisting of sampled values for the auxiliary variables in $\mathcal{D}_s$ based on their priors, 30 sets of observations $\epsilon_1, \ldots, \epsilon_{30}$ corresponding to 30 estimation events and the true state of the $Class$ variable, needed for testing.

Discrete variables O and R were associated with conditional probability tables $P(O|Class, \Delta\pi(O))$ and

$P(R|class, \Delta pa(R))$ that encoded noisy observations, such that the classification error rates based on O and R only, exceeded 25%, without introducing any shifts. In other words, in the experiments observations of A, B and C contributed significant information for the overall classification. In all experiments, variables A, B, C, O and R were observed, while variable E was latent.

Also, in the experiments the distributions over O and R were not subject to distribution shifts, i.e. the states of $\Delta\pi(O)$ and $\Delta\pi(R)$ were instantiated to zero shifts during the data sampling process. The same was true for $Class$. However, in all experiments the full extended model was used for the shift estimation , in which also $\Delta\pi(O)$, $\Delta\pi(R)$ and $\Delta\pi(Class)$ had to be estimated, reflecting the fact that no prior knowledge about any shift was made, except the form of distributions.

In the first series of experiments the states of auxiliary variables $\Delta\pi(A)$, $\Delta\pi(B)$ and $\Delta\pi(C)$ were randomly sampled along with the observations of A, B, C, R, O and the true class label of the last event. In this way distribution shifts in the modelling components representing sensors A, B and C were introduced. The sensor variance was kept low, as this introduced greater discrepancies between $P(\mathcal{V})$ and $P(\mathcal{V})^*$. Moreover, the shifts $\Delta\pi(E)$, $\Delta\pi(O)$, $\Delta\pi(R)$ and $\Delta\pi(Class)$ were set to zero. The first row in Table I shows the 27.5 % classification error rate of the original model trained on the data sampled from the initial distribution $P(\mathcal{V})$ but used for the classification of observations ϵ sampled from $P(\mathcal{V})^*$. The structure of this model is shown in Fig. 1. Subsequent rows in Table I show the classification error rates achieved by using different numbers of estimation events. The second row shows 14.6 % classification error rate achieved on the same data set using the extended model corresponding to the graph in Fig. 3, i.e. using the observations from a single event. As the number of used estimation events grows, the error rates drop to 4.6% for 30 events.

In the second series of experiments states of $\Delta\pi(E)$ were randomly sampled along with the observations of A, B, C, R, O and the true class label of the last event. Moreover, the shifts $\Delta\pi(A)$, $\Delta\pi(B)$, $\Delta\pi(C)$, $\Delta\pi(O)$, $\Delta\pi(R)$ and $\Delta\pi(Class)$ were set to zero during data sampling. The first row in Table II shows the classification error rate 34.8% of the original model trained on the data sampled from the initial distribution $P(\mathcal{V})$. Subsequent rows in Table II show the classification error rates achieved by using different numbers of estimation events. The second row shows 16.33 % classification error rate achieved on the same data set using the extended model corresponding to the graph in Fig. 3 , i.e. using observations from a single event. As the number of estimation events reaches 30, the error rate drops to 9.6%.

In the third series of experiments $\Delta\pi(A)$, $\Delta\pi(B)$, $\Delta\pi(C)$ and $\Delta\pi(E)$ were randomly sampled along with the observations of A, B, C, R, O and the true class label of the last event. The first row in Table III shows the classification error rate 40.2% of the original model trained on the data sampled from the initial distribution $P(\mathcal{V})$. Subsequent rows in Table II show the classification error rates achieved by using different numbers of estimation events. The second row shows 17.5 % classification error rate achieved on the same data set using the extended model corresponding to the graph in Fig. 3, an error reduction by more than 56 %. As the number of used estimation events reaches 30, the error rate drops to 11.1% .

Experiment	Classification Error
Don't adapt & test on $P(\mathcal{V})^*$	27.5%
Auto adapt 1 & test on $P(\mathcal{V})^*$	14.6%
Auto adapt 2 & test on $P(\mathcal{V})^*$	10.8%
Auto adapt 3 & test on $P(\mathcal{V})^*$	8%
Auto adapt 4 & test on $P(\mathcal{V})^*$	6.2%
Auto adapt 30 & test on $P(\mathcal{V})^*$	4.6%

Table I: Experiments with random distribution shift of the conditional probabilities of A, B and C.

Experiment	Classification Error
Don't adapt & test on $P(\mathcal{V})^*$	34.8%
Auto adapt 1 & test on $P(\mathcal{V})^*$	16.33%
Auto adapt 2 & test on $P(\mathcal{V})^*$	14 %
Auto adapt 3 & test on $P(\mathcal{V})^*$	12.6%
Auto adapt 4 & test on $P(\mathcal{V})^*$	11.9%
Auto adapt 30 & test on $P(\mathcal{V})^*$	9.6%

Table II: Experiments with random distribution shifts of the conditional probabilities of E .

Finally, the extended models along with the self adaptation approach were also tested on the data sampled from $P(\mathcal{V})$ prior to the distribution shifts. In this case it is expected that the self adaptive approach cannot achieve the same classification accuracy as the original model assuming zero distribution shifts (i.e. a model corresponding to the graph in Fig. 1. Table IV shows classification error rates for multiple cases: row 1 shows the best possible classification accuracy achieved by the ground truth model representing $P(\mathcal{V})$; in row 2 to 5 show the classification error rates for different numbers of estimation events considered for the estimation of distribution shifts. The error rates dropped from 8.5 % to 1 % as the number of events is increased from 1 to 30.

A. Discussion of the Results

The experimental results confirmed the expected impact of distribution shifts and the potentially useful properties of the presented method.

Firstly, using original models on data sampled from distributions that have undergone significant shifts can result in

Experiment	Classification Error
Don't adapt & test on $P(\mathcal{V})^*$	40.2%
Auto adapt 1 & test on $P(\mathcal{V})^*$	17.5 %
Auto adapt 2 & test on $P(\mathcal{V})^*$	15.2%
Auto adapt 3 & test on $P(\mathcal{V})^*$	14.2 %
Auto adapt 4 & test on $P(\mathcal{V})^*$	13.7%
Auto adapt 30 & test on $P(\mathcal{V})^*$	11.1%

Table III: Experiments with random distribution shifts of the conditional probabilities of A, B, C and E.

Experiment	Classification Error
Known $P(\mathcal{V})$ & test on $P(\mathcal{V})$	0.7%
Auto adapt 1 & test on $P(\mathcal{V})$	8.47 %
Auto adapt 2 & test on $P(\mathcal{V})$	7.33%
Auto adapt 3 & test on $P(\mathcal{V})$	6.1 %
Auto adapt 4 & test on $P(\mathcal{V})$	4.7%
Auto adapt 30 & test on $P(\mathcal{V})$	1%

Table IV: Experiments with no distribution shifts.

large increase of error rates. This is clearly visible in the first row of Tables I, II and III, where the error rates increased by more than an order of magnitude compared to the expected performance of the optimal model trained and tested on the original distribution $P(\mathcal{V})$.

Secondly, the extended model in combination with inference did introduce inherent robustness. Already the extended model for one estimation event significantly reduced the impact of the distribution shifts, as seen on the second row of Tables I, II and III.

Thirdly, the models representing multiple estimation events significantly reduced the error rates by correlating observations from subsequent events. In fact, the estimation of distribution shifts appears to be effective already with a few estimation events. While this seems to be a consequence of relatively simple parametric distributions used in the experiments, it is still surprising. This is likely to be a consequence of the inference exploiting heterogeneous observations from each estimation event. Clearly, the more complex parametric distributions are used, shifts for more parameters have to be estimated, requiring reasoning over more estimation events.

Also,the experiments confirm that the method supports estimation o distribution shifts related to different parts of the models, including the parameters of latent variables. Clearly, the effectiveness of the method depends also on the redundancy of the data source as well as the dependencies in the actual probability distributions from which the training and operational data were sampled.

Finally, Table IV shows the impact of the automatic shift estimation on the model performance in the case no shifts were introduced, i.e. the model was trained and tested on the data sampled from the same distribution $P(\mathcal{V})$. The extended models do introduce additional uncertainty to inference resulting in error rates that are for larger numbers of estimation events slightly higher than the error rates using the optimal model trained and tested on the initial $P(\mathcal{V})$ data distribution. For example, for 30 events the presented approach achieved 1.% error rate vs. 0.7 % error rate achieved by the ground truth model. For smaller numbers of events, the presented approach produces higher error rates, such as 8.5.% for one event, 6.1.% for three events, etc. However, the performance of the initial model that was not adaptive dropped significantly after the introduction of the distribution shifts resulting in $P(\mathcal{V})^*$, whereas the adaptive models continued performing with low error rates prior and after the distribution shifts. For example, as indicated by Tables I and IV, the adaptive model of four estimation events had error rates of 4.7% on data sampled from the original distribution $P(\mathcal{V})$ while it had the error rate of 6.2 % on the data from $P(\mathcal{V})^*$ that was obtained by shifting distributions of the observation models associated with nodes A, B and C. The optimal model without auxiliary shift variables, on the other hand, achieved 0.7 % error rates on $P(\mathcal{V})$, but was significantly worse on the data sampled from $P(\mathcal{V})^*$, where the error rates exceeded 27%. Similar tendencies can be observed by inspecting Tables II and III.

In other words, we sacrifice some classification accuracy when there is no shift for the overall robustness of the model to distribution shifts.

VI. Applications

The observed properties can be exploited in multiple ways, especially if Probabilistic Graphical Models are used.

Firstly, the auxiliary variables can be directly integrated into the model supporting the enhanced inference processes. This can result in significantly improved robustness against distribution shifts and even in automatic adaptation of the modelling components, such that the distribution shifts are mitigated.

Moreover, such extended models could be used also next to the classifiers, supporting auxiliary monitoring of the distribution over relevant variables. The inference on such models using the same data as the main classifiers could be used to indicate potential modelling discrepancies, providing a warning about potential degradation of the main classifier. If the main classifiers used the same models of the correlations between the observable variables and the class variable, then the inference on each auxiliary shift variable would directly pinpoint the main modelling components that should be adapted. For example, if inference on the extended model shown in Fig. 3 yielded a posterior $P(\Delta A|\epsilon)$ indicating a significant distribution shift, the CPT $P(A|E)$ in the original model should be updated. Alternatively, the CPT $P(A|E)$ pinpointed by the inference process could simply be deactivated or removed from the model, which is eliminates the impact of the shift.

VII. Conclusions

Modern AI based solutions increasingly rely on statistical domain models that capture correlations between variables of interest. Such models in combination with inference algorithms support reasoning about data patterns obtained during operation. Statistical models are typically obtained through machine learning using training data sampled under certain conditions in the domain. Unfortunately, the data during operation is often sampled from a different distribution than the data that was used for the training of the models. Such distribution shifts can have a significant impact of the quality of the inference processes which, in turn, can have adverse effects on the application functions. Unfortunately, distribution shifts can often not be avoided.

The proposed approach uses probabilistic modelling and inference to (i) identify distribution shifts, (ii) introduce inherent

robustness against distribution shifts and (iii) automatically adapt the models to identified distribution shifts.

The approach assumes probabilistic reasoning over sets of dependent variables representing heterogeneous sources of data. The modelling and inference methods introduced by probabilistic graphical models enable systematic extension of probabilistic domain models with auxiliary variables supporting explicit probabilistic inference about possible distribution shifts. In this way, the distribution shifts associated with different components of the model can be identified. Modelling over multiple events enables increasingly accurate estimation of the distribution shifts. The estimation results can contribute to improved trust in a system in multiple ways. Firstly, they can be seen as a "data shift gauge" measuring the shift and informing the users about situations in which the solution's performance might have degraded. If the distribution shifts in the true data generating processes are limited to small subsets of variables and occur over extended periods of time, the identified modelling components could be repaired with data sets that are small compared to the sets required for relearning of the entire model. Secondly, the model extensions are likely to improve the inference robustness against distribution shifts. This is possible as the model extensions allow explanations of combinations of the observed data that were not available during training. Finally, by exploiting the inference about auxiliary variables over multiple estimation events, the impact of the distribution shifts can be to a great extent reduced/eliminated.

These properties were confirmed in controlled experiments with synthetic data. By manipulating ground truth models, distribution shifts corresponding to different components of the models were introduced in a systematic way. While the experimental results reported in the paper correspond to a specific ground truth model, the same tendencies were observed in experiments with different models. Clearly, the effectiveness of the approach depends on the actual domain, the sets of observed variables and dependencies between them. The more variables representing heterogeneous sources and the sparser dependencies between them, the greater is effectiveness of the presented solution. The paper used a simple basic model with a tree topology for the sake of clarity. However, the presented method can be applied to complex models consisting of multiple loops. Also, a single estimation event could correspond to a dynamic process represented with a Dynamic Bayesian Network, instead of a static model in Figure 1. Also, in this paper it was assumed that the distribution shifts take place over longer periods of time, such that they can seen as constant during a multi-event estimation process.

Future work will focus on applying the presented method to real world cases and efficient recursive estimation of distribution shifts over time.

REFERENCES

[1] G. Pavlin, A.-L. Jousselme, J. P. de Villiers, P. C. Costa, K. Laskey, F. Mignet, and A. de Waal, "Online system evaluation and learning of data source models: a probabilistic generative approach," in *2019 22th International Conference on Information Fusion (FUSION)*, 2019, pp. 1–10.

[2] R. Alaiz-Rodríguez and N. Japkowicz, "Assessing the impact of changing environments on classifier performance," in *Conference of the Canadian Society for Computational Studies of Intelligence*. Springer, 2008, pp. 13–24.

[3] D. A. Cieslak and N. V. Chawla, "A framework for monitoring classifiers' performance: when and why failure occurs?" *Knowledge and Information Systems*, vol. 18, no. 1, pp. 83–108, 2009.

[4] J. Quiñonero-Candela, M. Sugiyama, A. Schwaighofer, and N. D. Lawrence, *Dataset shift in machine learning*. Mit Press, 2008.

[5] D. J. Hand, "Classifier technology and the illusion of progress," *Statistical science*, vol. 21, no. 1, pp. 1–14, 2006.

[6] J. G. Moreno-Torres, T. Raeder, R. Alaiz-Rodríguez, N. V. Chawla, and F. Herrera, "A unifying view on dataset shift in classification," *Pattern recognition*, vol. 45, no. 1, pp. 521–530, 2012.

[7] M. J. Roemer and G. J. Kacprzynski, "Advanced diagnostics and prognostics for gas turbine engine risk assessment," in *2000 IEEE Aerospace Conference. Proceedings (Cat. No. 00TH8484)*, vol. 6. IEEE, 2000, pp. 345–353.

[8] X. Li, W. Zhang, H. Ma, Z. Luo, and X. Li, "Deep learning-based adversarial multi-classifier optimization for cross-domain machinery fault diagnostics," *Journal of Manufacturing Systems*, vol. 55, pp. 334–347, 2020.

[9] M. Higger, "Robust fusion methods for distribution shifts," Ph.D. dissertation, Northeastern University, 2013.

[10] P. Costa, K. Laskey, E. Blasch, and A.-L. Jousselme, "Towards unbiased evaluation of uncertainty reasoning: The urref ontology," in *Proceedings of the 15th International Conference on Information Fusion*, Singapore, July 2012.

[11] "Evaluation of Techniques for Uncertainty Representation Working Group Website." [Online]. Available: http://eturwg.c4i.gmu.edu/

[12] J. P. de Villiers, K. Laskey, A.-L. Jousselme, E. Blasch, A. de Waal, G. Pavlin, and P. Costa, "Uncertainty representation, quantification and evaluation for data and information fusion," in *18th Int. Conf. on Information Fusion (Fusion)*, July 2015, pp. 50–57.

[13] J. P. de Villiers, A.-L. Jousselme, A. de Waal, G. Pavlin, K. Laskey, E. Blasch, and P. Costa, "Uncertainty evaluation of data and information fusion within the context of the decision loop," in *19th Int. Conf. on Information Fusion (FUSION)*, July 2016, pp. 766–773.

[14] J. P. de Villiers, R. W. Focke, G. Pavlin, A.-L. Jousselme, V. Dragos, K. B. Laskey, P. C. Costa, and E. Blasch, "Evaluation metrics for the practical application of URREF ontology: An illustration on data criteria," in *20th Int. Conf. on Information Fusion (Fusion)*, July 2017.

[15] B. Lewis, T. Scarnati, E. Sudkamp, J. Nehrbass, S. Rosencrantz, and E. Zelnio, "A sar dataset for atr development: the synthetic and measured paired labeled experiment (sample)," in *Algorithms for Synthetic Aperture Radar Imagery XXVI*, vol. 10987. International Society for Optics and Photonics, 2019, p. 109870H.

[16] E. Blasch, U. Majumder, E. Zelnio, and V. Velten, "Review of recent advances in ai/ml using the mstar data," in *Algorithms for Synthetic Aperture Radar Imagery XXVII*, vol. 11393. International Society for Optics and Photonics, 2020, p. 113930C.

[17] N. Inkawhich, M. J. Inkawhich, E. K. Davis, U. K. Majumder, E. Tripp, C. Capraro, and Y. Chen, "Bridging a gap in sar-atr: Training on fully synthetic and testing on measured data," *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, vol. 14, pp. 2942–2955, 2021.

[18] E. Blasch, J. Bin, and Z. Liu, "Certifiable artificial intelligence through data fusion," *arXiv preprint arXiv:2111.02001*, 2021.

[19] G. Pavlin, A. Jousselme, J. P. de Villiers, P. C. G. Costa, and P. de Oude, "Towards the rational development and evaluation of complex fusion systems: A urref-driven approach," in *21st International Conference on Information Fusion, FUSION 2018, Cambridge, UK, July 10-13, 2018*. IEEE, 2018, pp. 679–687. [Online]. Available: https://doi.org/10.23919/ICIF.2018.8455719

[20] F. V. Jensen and T. D. Nielsen, *Bayesian Networks and Decision Graphs*, 2nd ed. Springer-Verlag, 2007.