How to Cope with Change?
Preserving Validity of Predictive Services over Time

Lucas Baier
KIT, Germany
lucas.baier@kit.edu

Niklas Kühl
KIT, Germany
kuehl@kit.edu

Gerhard Satzger
KIT, Germany
gerhard.satzger@kit.edu

Abstract

Companies more and more rely on predictive services which are constantly monitoring and analyzing the available data streams for better service offerings. However, sudden or incremental changes in those streams are a challenge for the validity and proper functionality of the predictive service over time. We develop a framework which allows to characterize and differentiate predictive services with regard to their ongoing validity. Furthermore, this work proposes a research agenda of worthwhile research topics to improve the long-term validity of predictive services. In our work, we especially focus on different scenarios of true label availability for predictive services as well as the integration of expert knowledge. With these insights at hand, we lay an important foundation for future research in the field of valid predictive services.

1. Introduction

Due to the large increase of data in recent years, various industries are trying to reap the benefits of this new resource for their service offerings. Machine learning is playing an important role in nearly all fields of business, ranging from marketing over governmental tasks to scientific-, health- and security-related applications [1]. Many companies rely on machine learning models deployed in their information systems for increasing the efficiency of their processes [2] or for offering new services [3]. As Davenport [4] describes, companies which are able to leverage their data sources through analytical tools achieve a substantial competitive advantage.

However, it is worth regarding how such predictive services based on machine learning are built, deployed and executed in the long run. Traditionally, supervised machine learning models are trained using historical data containing input features and a corresponding target [5]. Subsequently, the model is used to continuously make predictions for a specific service (e.g. the failure of a machine) on a stream of unseen incoming data. We define such a service as a "predictive service". However, data streams typically evolve over time and thus, their data structure or the underlying probability distribution changes [6]. This depicts a challenge since supervised machine learning models are very sensitive to changes in their input data, e.g. to the adjustment of production parameters [7]. Even small deviations can have significant impact on the deployed model—drastically influencing its prediction performance and the utility of the predictive service [8]. However, it is difficult to detect this change in the input data and, furthermore, to adapt the model accordingly [7]. In the field of computer science, the phenomenon of a changing relation over time between the input features and the target labels is predominantly called "concept drift" [9].

An example for an application with evolving data over time is a predictive service which monitors the output quality in a chemical production process and predicts corresponding failures [10]. Such a predictive service relies on the input data generated by the sensors that the production machine is equipped with. Sensors wear out over time [11] and the resulting measurements change accordingly, leading to different input data. However—without the necessary precautions—a machine learning model is not prepared for this change since this pattern has not been observed before in the training set. Thus, meaningful quality predictions are impossible to make in the long run, and the service does not keep up to its promised validity. Therefore, we define a general research question which guides this research paper:

General RQ. *How can we design an effective and efficient automated artifact for predictive services, which ensures their long-term validity?*

Based on this general research question, we aim to describe the current status of predictive services.

RQ1. *How can we distinguish between various forms of existing predictive services with regard to their lifecycle?*

For answering this question, this work introduces a definition of predictive services as well as a

URI: https://hdl.handle.net/10125/59548
ISBN: 978-0-9981331-2-6

framework for characterizing predictive services with respect to their validity over time. The framework can be used as a support tool for practitioners during the introduction of a predictive service so that all relevant design options are considered. Furthermore, the framework allows a thorough analysis as well as comparison of existing approaches. In demarcation to existing frameworks such as Gama et al. [9], our framework includes the setup as well as operation phase of predictive services. Subsequently, we classify available research papers into the framework resulting in a heatmap which serves as a foundation for deriving a research agenda of valuable research topics: The different availability of true labels during operation as well as methods for domain expert integration.

The remainder of the paper is structured as follows: Section 2 presents related work on which we base our research and introduces a definition of predictive services. Section 3 presents a framework for characterizing aspects of predictive services and classifies existing practical research on that basis. Section 4 introduces research opportunities that are derived from challenges identified in the previous section. The fifth and final section discusses our results, describes theoretical and managerial implications, acknowledges limitations and outlines future research.

2. Foundations

To allow for a common understanding, we first introduce the theoretical foundations for the examination of the validity of predictive services. We give a brief overview on machine learning for services, followed by an overview of research in computer science that deals with concept drift. Subsequently, we introduce predictive services.

2.1. Machine learning for services

Machine learning in general has recently received a lot of attention due to the massive flow of available data and increasing computation power. Traditionally, approaches are divided into supervised and unsupervised machine learning [12]. Supervised machine learning depends on labeled examples in the training data, whereas unsupervised machine learning aims at detecting unknown patterns in the data. Most real-word applications of machine learning are of supervised nature ([13], [14]). Therefore, we focus on supervised approaches in the following. Well-known application examples are the prediction of a credit rating or the fingerprint matching on current smartphones.

The importance of analytical and machine learning solutions for service science has been highlighted by the introduction of service analytics [15]. Service analytics describes the dedicated application of analytical tools such as machine learning on data created in service systems to improve or extend existing service offerings. In this context, continuous data streams over time play an important role. Machine learning is, for instance, applied to monitor click streams on web pages or to monitor events and notifications [16]. All those examples are confronted with changing data streams over time. Therefore, the next section introduces definitions as well as solutions developed for this challenge from a computer science perspective.

2.2. Concept drift

The computer science community has examined the challenge of changing data streams in machine learning over time under the term "concept drift" [17]. A concept $p(X, y)$ is described as the joint probability distribution over a set of input variables X and the label or target variable y. However, "in the real world concepts are often not stable but change with time" [7, p. 1]. This leads to the problem that machine learning models built on previous data are not valid anymore for new incoming data which requires regular model updates or retraining. There exists a variety of descriptive definitions of concept drift ([6], [13], [14]). A mathematical definition is given by Gama et al. [9]:

$$p_{t0}(X, y) \neq p_{t1}(X, y)$$

The definition states that we are facing a concept drift if there is a difference in the concept at *t0* compared to the concept at *t1*. This change of the joint distribution is challenging for supervised machine learning models since they are typically trained on a fixed initial training set [7]. However, if the features and the label of the training set just belong to the concept at *t0*, the model is only trained to recognize objects of the first concept whereas it does not know how to handle instances belonging to the second concept at *t1*. Changes in the incoming data stream can depend on many internal or external factors. Therefore, it is intuitive that different types of changes in data streams can be identified. One popular classification of concept drift depicts four different types [19]: Sudden concept drift, incremental concept drift, gradual concept drift and reoccurring concept drift such as seasonal patterns. Webb et al. [20] also provide a more detailed taxonomy with categories such as drift and concept duration as well as drift magnitude.

Gama et al. [9] introduce a framework which focuses on algorithmic methods for changing data

streams. The framework consists of four categories: Memory, change detection, learning and loss estimation. A description on application-related use cases is given by Žliobaite et al. [10]. They provide a list of 54 research papers that implement solutions and methods for changing data streams with real data. Concrete use cases which consider the challenges of concept drift occurrence can be clustered into monitoring and control tasks ([17], [18]), information management ([19], [20]) and analytics and diagnostics tasks ([21], [22]).

Based on the foundations in the previous two subsections, we introduce predictive services in the following.

2.3. Predictive services

We define predictive services as services based on predictions that are acquired through the application of supervised machine learning models on data available in its service system environment. Predictive services are fully deployed on a productive IT infrastructure and thereby are constantly issuing new predictions. The final objective can either be the delivery of the prediction itself (e.g. forecast for the market demand for a product) or an action based on the prediction (e.g. the automated adjustment of the production schedule for a product).

We assume that the validity of predictive services can be affected in two ways: First, the *environment of the service* changes, which influences the resulting data, and thus, the quality of the prediction. This is the case when a sensor on a production machine wears out over time and delivers less reliable results. Second, the *application of the service* itself affects its predictive power over time. The second case can be illustrated by a predictive policing service indicating the neighborhoods in a city with the most criminal activities. The local police will accordingly reinforce their presence in this area which results in a decreasing criminal statistic over time. This development, however, will invalidate the recommendations of the predictive service which continues to classify this neighborhood as a high-risk area [26]. After all, any kind of predictive service is facing the challenge of a changing environment over time; it is just a matter of the time span that is considered.

The example above illustrates the complexity of ensuring the long-term validity of predictive services. Therefore, the problem requires a comprehensive and interdisciplinary analysis. On the one hand, it is necessary to thoroughly examine the technological side of the problem. On the other hand, the economic side must be also considered, and benefits or downsides of possible solutions must be assessed.

The next section introduces a framework which can be used to set up a predictive service and to prepare it for changes in the data stream in order to guarantee the validity over time.

3. Conceptual framework

The framework can be understood as a tool to support the initiation and implementation of a predictive service. It gives guidance for decisions during the setup phase of a predictive service but also provides solutions for challenges during the operation and use of the predictive service. Furthermore, it allows to differentiate between characteristics of predictive services. At first, we explain the methodology that we applied for the development of the framework. Subsequently, we introduce the framework itself which is split into three parts. The first part relates to necessary design decisions during the setup of a predictive service. The second part displays the algorithmic options for keeping the validity over time. The third part presents characteristics that need to be considered during operation, especially the availability of true labels and the constant evaluation of the service.

3.1. Methodology

Gama et al. [9] provide a taxonomy which explains the different algorithmic options for handling changing data streams. This taxonomy is the basis for our framework and is mainly reflected in the second part (c.f. section 3.3). However, their taxonomy is missing design decisions during setup as well as operation of predictive services. The consideration of both phases is (besides the algorithmic methods) crucial for the development of a successful predictive service. Our framework is therefore built as an extension to the prevailing taxonomy.

We developed our framework by a rigorous analysis of existing use cases in research that examine concept drift. We base this analysis on the 54 research papers which are presented in Žliobaite et al. [10], as those include papers from a wide range of application tasks. We remove all research papers with unsupervised approaches and those that do not provide sufficient information for in-depth comparisons resulting in 23 remaining research papers. Based on a forward and backward search on this list, we identified 11 additional research papers. In total, we included 34 research papers ([11], [21], [23], [25], [27-56]) into our detailed analysis. During the analysis, we iteratively added or removed categories as we progressed with the number of research papers. The items are based on existing literature. If we could not

identify suitable literature, we added the items based on our analysis of the research papers. The resulting framework needs to be understood as an exploratory tool which still develops over time as new research papers are included.

3.2. Setup decisions for predictive service

Before a predictive service can be offered, fundamental decisions about the setup of the service must be made. Table 1 depicts the different categories for the setup phase.

Business focus: When designing a predictive service, one of the first steps is to clearly define the business focus. What is the benefit that the service is delivering and who is the final user/customer of it? The customer can either be external (e.g. a service provider offers constant social media analytics to a customer with a tool) or internal (e.g. predictive service is used for the improvement of internal processes) [2].

Table 1. Predictive service setup decisions

Business focus	External	Internal	Unknown		
Data input	Structured	Unstructured			
Machine learning task	Regression	Classification			
Domain expert knowledge	Label provision	Feature generation	Model building	Change detection	None
Type of change	Sudden / Abrupt	Incremental / Gradual	Reoccurring	Unknown	

Data input: A differentiation with regard to the data input which is used for the predictive service is necessary. Structured data in form of tables (e.g. [38]) can easily be utilized by most machine learning algorithms and change detection approaches. However, unstructured data (e.g. text data [51]) is more complex to process and requires more advanced handling techniques, especially for the change detection.

Machine Learning Task: A clear definition of the relevant machine learning task behind the predictive service is indispensable. If the aim is to predict the continuous value of a target, regression techniques have to be applied (e.g. [21]). If the aim is to predict the class membership of an object, classification will be used (e.g. [28]).

Inclusion of domain expert knowledge: The knowledge of domain experts is a valuable resource for the validity of predictive services [10]. Several ways in which domain experts can support the development of valid predictive services have been identified. The simplest way to include domain experts into the process is the provision of true labels for the service. For instance, a predictive service is monitoring the quality in a chemical production facility. True labels for the chemical product can be acquired from experts who examine selected samples in a laboratory. Domain experts can also be included into the feature generation process for the machine learning model [57]. Especially experienced machine operators often know which hints and signals are relevant for the prediction of a machine failure and jointly it can be thought how to transform this information into a feature for the learning algorithm. It is also possible to explicitly apply knowledge of domain experts during the model building process, e.g. through the inclusion of fixed decision rules. Domain experts can also be relevant for the explicit detection of changes in the data. Human experts supported and empowered by advanced visual analytics tools can provide more insights than an algorithm alone [58].

Type of change: During the setup phase of a predictive service, expected changes of the data stream which affect the validity can already be identified. If this information is known a priori, suitable algorithms can be chosen beforehand. The different types of changes are based on the definition by Žliobaite [19]. Sudden concept drift refers to situation where the data changes abruptly from one time point to another. Incremental and gradual concept drift both refer to situations where the change in the data stream happens slower over time. The two types are merged here since in real use cases the two are mainly not differentiable. Reoccurring concept drift refers to situations where data changes regularly to already known patterns such as seasonal contexts.

3.3. Algorithmic decisions

The second part of the framework relates to the algorithmic and technical characteristics of the predictive service. This subsection is built on the research paper by Gama et al. [9] which identifies four categories for dealing with changing data over time:

Memory, Detection model, Learning, Loss estimation. The items for each category are also based on the work by Gama et al. [9], however their item specification is very detailed. During our analysis, we realized that items can be merged without information loss. Table 2 contains the corresponding categories as well as the items that we specified during our analysis.

Memory: Due to the massive amount of data produced in data streams, it is often infeasible to consider all data instances of a data stream. This category deals with the memory management of the predictive service. How many instances are stored for training or retraining of the algorithm? The quantity can range from a single instance to multiple or all instances. Often, algorithms only consider a window of the last *n* instances which are deemed to be still relevant to the algorithm. In cases with massive computing power or limited size of data in the stream, the algorithm might also consider all instances. It is also possible that only a sample of past data is used.

Table 2. Algorithmic decisions for predictive services

Memory	Single	Multiple (window)	All (gradual forgetting)	All (no forgetting)	Sampling
Detection model	Sequential analysis	Control chart	Two distri-butions	Contextual	Others
Learning mode	Retraining + single	Incremental + single	Retraining + ensemble	Incremental + ensemble	
Loss estimation	Model independent		Model dependent		

Change detection: Change detection refers to the mechanism that is applied to detect a change in the data stream. Various approaches have been proposed in research. In sequential analysis, the values of new data instances are compared to older values on the basis of statistical tests. Other approaches rely on statistical process control which is widely applied in chemical production processes. The algorithm tracks the number of correct predictions over time and if the amount of false predictions exceeds a predefined threshold, an alarm is triggered. However, this approach requires the instant provisioning of the true labels after the prediction. Another way is the application of two time-windows with different size. The statistical data distributions of the two windows are compared with statistical tests. In case of a difference, a change or concept drift has happened. Contextual approaches use time-related measures for change detection.

Learning: As soon as new true labels for previous predictions are available to the predictive service, the machine learning algorithm behind it might be adapted. Usually, two different options are available: Retraining, where the old model is discarded and a new one is trained from the scratch or incremental updates, where the current model is slightly modified. Incremental learning is closely connected to the idea of continuous learning where the model never stops to learn according to the circumstances. Concerning the type of model, it can be differentiated between a single model or ensemble models where several models are combined for a prediction.

Loss estimation: Supervised machine learning models rely on feedback/true labels to optimize their performance. One can differentiate between model-dependent and model-independent loss estimation methods. Model-independent loss estimation approaches are more popular where a metric such as accuracy is computed and evaluated over time. However, some machine learning techniques such as Support Vector Machines allow the detection of changes in the data based on internal algorithmic characteristics.

While we now discussed the necessary characteristics of the setup of valid predictive services, the next section describes challenges during the operation of predictive services.

3.4. Operation of predictive service

During the operation of a predictive service, constant updates and improvements are necessary. Therefore, relevant topics are the acquisition of true labels as well as the evaluation criteria as depicted in table 3.

Label: The availability of true labels during operation is the most relevant feedback for the optimization of a machine learning algorithm deployed on a data stream. Therefore, this category is highly important to guarantee the validity and proper functionality of predictive services. Label availability is differentiated into three items: Full label, limited label and no label availability.

Full label availability refers to the case where the predictive service can receive access to all true labels after the prediction. Thus, the service receives feedback to every single prediction that it issued before, and the algorithm constantly receives new training data for improvement. Weather predictions are an example for this item. If the service issues a

weather prediction for the next day, we can always receive the true label for the weather on the following day—and continue to learn on these insights.

Table 3. Operation of predictive service

Label availability	Full	Limited	None
Evaluation criteria	Statistical evaluation metrics	Statistical evaluation combined with business impact	

Full label availability refers to the case where the predictive service can receive access to all true labels after the prediction. Thus, the service receives feedback to every single prediction that it issued before, and the algorithm constantly receives new training data for improvement. Weather predictions are an example for this item. If the service issues a weather prediction for the next day, we can always receive the true label for the weather on the following day—and continue to learn on these insights.

Limited label availability means that only a fraction of all true labels can be accessed after the prediction. In this case, the algorithm only receives feedback on its performance for a few instances. A further differentiation can be made by determining whether it is possible to select the instances for which labels are acquired (e.g. true quality of a specific chemical product can be determined by a laboratory analysis) or whether it is a random sample. An example for this is a predictive service determining customer satisfaction and true labels are received by sending a survey to all customers. However, we do not know who is going to respond to the inquiry. Therefore, the instances in the sample cannot be influenced and are random.

No label availability describes a situation when it is impossible to acquire labels. During training of the prediction model, a full data set with labels is available. However, during operation, when the predictive service is deployed, no true labels for previous predictions can be received. Therefore, the machine learning model cannot adapt its predictions to changes in the data. This demands methods that are specifically robust to outliers and unexpected deviations in the data [8]. Reasons for no label availability can be that it is too costly to acquire the true labels. In other situations, it might just be impossible to receive the true labels, e.g. a machine part for whose functionality we can receive true labels with sensors in a specialized test bench; however, in the field of application these sensors are not available and therefore labels are impossible to derive.

Evaluation criteria: The traditional evaluation of the performance of machine learning models is based on statistical evaluation metrics such as accuracy, recall or F1-score [12]. These metrics are suitable for expressing the mere algorithmic performance on the use case that is reflected. However, since this work considers the explicit service based on the algorithm, it is also necessary to study the business impact of the predictive service, especially the influence of validity over time [9]. One way is to examine the influence on profits. Many use cases where predictive services are applied also lead to imbalanced cost of prediction mistakes. In case of predictive maintenance, it is costlier to not predict and therefore miss the failure of a machine resulting in a very expensive stop of the whole production instead of triggering a false alarm. It is also necessary to consider the environment where the predictive service is deployed. This refers to computational but also memory constraints in the IT infrastructure. Investment and setup costs also need to be considered. This category is closely linked to the business focus category in section 3.2.

3.5. Heatmap of research papers

In the following paragraph, we classify the 34 research papers which we used for the development of the framework. The result of this approach is a heatmap which is depicted in table 4.

Many application cases utilize several of the design options in parallel or test different variations in their approaches. Therefore, the sum of papers per row often exceeds 34. The heatmap indicates the different design options which were chosen by the different researchers. This allows to understand which of the available solutions and methods are really implemented for use cases and how often they are used. As stated above (section 3.1), the heatmap has to be understood as an exploratory tool since we do not map all existing research papers.

The heatmap indicates that current use cases dealing with changing data over time mainly use structured data for a classification problem with sudden or incremental changes in the data (e.g. [23], [28], [31]). There seems to be a lack in the consideration of economic challenges. Many projects do not name a specific business focus behind the implemented prediction model (e.g. [20]). The reason for this may lie in the academic nature of the projects. Furthermore, most use cases rely on statistical evaluation only (e.g. [11], [48]). However, this consideration lacks evidence whether its economically viable and useful to implement such a service.

Table 4. Heatmap of existing research classified into the framework

Bus. focus	External **4**	Internal **4**	Unknown **26**		
Data input	Structured **25**	Unstructured **9**			
Mach. learning	Regression **5**	Classification **29**			
Domain expert	Label provision **17**	Feature generation **18**	Model building **2**	Change detection **1**	None **5**
Type of change	Sudden / Abrupt **24**	Incremental / Gradual **31**	Reoccurring **2**	Unknown **0**	
Memory	Single **1**	Multiple (window) **25**	All (gradual forgetting) **9**	All (no forgetting) **4**	Sampling **3**
Detect. model	Sequential Analysis **1**	Control Chart **12**	Two distri-butions **3**	Contextual **5**	Others **14**
Learn. mode	Retraining + Single **15**	Incremental + Single **11**	Retraining + Ensemble **2**	Incremental + Ensemble **9**	
Loss estim.	Model independent **33**	Model dependent **1**			
Label avail.	Full **21**	Limited **13**	None **0**		
Eval. criteria	Statistical evaluation metrics **31**	Statistical evaluation combined with business impact **3**			

0 >0 & <5 ≥5 & <10 ≥10 & <20 >20

Additionally, so far, the knowledge of domain experts is mainly used for label provision and feature generation (e.g. [21], [29]). Efficient methods for expert integration into model building and change detection are missing. Most research projects also assume a full availability of true labels for the predictive service (e.g. [25], [32], [47]). Only few approaches have been developed for a limited label availability (e.g. [39]) and there is no approach in our paper selection which deals with no label availability. However, those two are the categories that prevail in real-world applications.

4. A research agenda for preserving validity of predictive services over time

The heatmap in the previous section indicates that there is still a lack of dedicated solutions for challenges during the design and operation of predictive services which remain valid over time. Based on our analysis, we identify two areas where current research approaches lack solutions so far.

RQ 2. *Which are suitable methods for ensuring the validity of predictive services with limited availability of true labels in operation?*

True labels for a prediction are a very relevant feedback mechanism for any kind of machine learning algorithm. However, for a predictive service in operation, this information is only partly available — if at all [59]. The proposed framework already depicts the different possibilities for the available number of labels. Additionally, Žliobaite et al. [10] define temporal dimensions when the true label is available to the predictive service. They differentiate this temporal dimension into real-time, time-lag and on demand. Real-time availability means that the labels are available in the next time period after the prediction. In other situations, true labels might arrive after a fixed or variable time lag. Asking a user for feedback is an example for a use case where the true labels can be acquired on demand. If we combine the temporal dimensions with the volume dimensions, several different scenarios emerge which are depicted in table 5.

Table 5. Different scenarios for label availability

Time / Volume	Next time period	Time-lag	On demand
Full	e.g. Klinkenberg et al. [29]	e.g. Black et al. [33]	e.g. Fdez-Riverola et al. [22]
Limited	?	?	?
None	? (no time differentiation)		

There exist various algorithms for predictive services with full label availability during operation. However, solutions for the other scenarios when only limited or no true labels are available to the predictive service are sparse so far. This is depicted by the question marks in table 5. RQ 2 aims at developing and establishing methods for each of the scenarios with a question mark. In case only a limited number of labels is available, it might be possible to derive the missing

labels with the help of the existing ones (e.g. in form of a semi-supervised approach [60]). Another approach might be an efficient method for the integration of expert knowledge which leads to the next research question.

RQ 3. *How can expert knowledge be leveraged to increase the long-term validity of predictive services?*

The knowledge of domain experts is a very valuable resource in any form of analytical solution. This research question deals with the challenge on how this expertise can be leveraged to increase the validity of predictive services. Therefore, this question aims at examining and evaluating methods for expert knowledge integration. Several areas for expert integration are already presented in the framework in section 3. With regard to label provision, it is interesting to examine which labelled instances are most useful for the predictive service in order to improve its importance. One possible solution could be the application of active learning [19], a machine learning technique. In case of changing data, the machine learning model asks for expert support in labeling the most important instances for ensuring its ongoing validity. This also relates to the limited label availability in RQ2.

Furthermore, a structured method to integrate experts into the model building process is necessary. Possible methods can be derived from approaches in other machine learning areas but also from research streams that already enabled the successful integration of expert knowledge, e.g. in decision support systems [18].

Basing a change detection algorithm on expert input requires a constant monitoring of the predictive service. However, for instance in most production plants, this is the case anyway. This setup allows to use the strengths of each player involved in this scenario. The algorithm can provide a constant monitoring and is not distracted by other activities. The human expert meanwhile can work on other tasks and is only alerted when unusual patterns are detected in the data. Supported by advanced visual analytics, the expert can then for instance identify the type of change that occurred in the data and act accordingly. Another approach is the inclusion of experts directly in the training phase of the prediction model. Domain experts can anticipate possible data drifts and a model can be tuned in order to detect these corresponding drifts.

Independently of the actual method that is applied, the development of an efficient integration method could also increase the acceptance and understanding of domain experts for automated decisions made by predictive services which is a common challenge in practice [9]. During the answering of the RQs, a strong focus should lie on the economic evaluation of the proposed solution. Resulting costs (e.g. setup costs, computational costs during operation) need to be rigorously compared to the economic consequences of fewer false predictions for the predictive service.

5. Conclusion

Companies are increasingly dependent on data for the offering of their services. Predictive services, which are services based on predictions by supervised machine learning, are playing an important role in this context. These services constantly issue predictions over time which are an important decision support or might even act autonomously. Therefore, it is of high importance that predictive services work reliably. However, data streams constantly evolve and change over time and thereby challenge the proper functionality of the predictive service. This work proposes research areas to ensure the validity of predictive services over time. The contribution of this paper is threefold.

First, we provide a definition of predictive services and explain how their validity over time can be influenced by changing data. Second, based on previous research projects that are handling changing data streams, we develop a framework which gives guidance to practitioners but also to researchers for setting up a new predictive service. Furthermore, it allows to differentiate between existing predictive services. Third, after classifying the existing research approaches into the framework, we identified two areas for improvement: The label availability in operation as well as the integration of domain experts. Correspondingly, we developed a research agenda which aims at developing solutions for those challenges. The derived research agenda is of high importance to any endeavor dealing with predictive services. It is important that such services are resilient against changes in the incoming data streams.

Besides these contributions, this work has limitations. Validity is only one aspect of predictive services which needs to be examined. However, a holistic view on predictive services requires that also other aspects such as organizational challenges are considered. Companies need to ensure that they have the required resources such as a skilled workforce and IT infrastructure available. Furthermore, legal requirements are gaining more and more importance. The introduction of GDPR in Europe poses many challenges for most companies [61]. Predictive services often rely on personal data (e.g. the operators of a machine) or are based on IP-relevant data sources.

With regard to the developed framework, we are aware that the number of papers that we analyzed is

limited, and we do not claim to have included all relevant research papers. As new papers are added to the framework, it still might change and adopt. Since this is work is a research agenda, its content is rather conceptual and further quantitative evaluation of the problems stated is needed. By conducting expert interviews with practitioners, we plan to further refine the research demand and the possible solution space.

The use of predictive services in productive environments is only at the beginning of its development. In the future, more and more services will rely on automated decisions based on machine learning algorithms. Therefore, it is very worthwhile to investigate methods to guarantee the long-term validity of those services.

References

[1] H. Chen, R. Chiang, and V. Storey, "Business Intelligence and Analytics: From Big Data to Big Impact," *Mis Q.*, vol. 36, no. 4, pp. 1165–1188, 2012.

[2] R. Schüritz and G. Satzger, "Patterns of Data-Infused Business Model Innovation," in *CBI 2016*, 2016, pp. 133–142.

[3] V. Dinges, F. Urmetzer, V. Martinez, M. Zaki, and A. Neely, "The future of servitization: Technologies that will make a difference," *Cambridge Serv. Alliance Exec. Brief. Pap.*, 2015.

[4] T. H. Davenport, "Competing on analytics.," *Harv. Bus. Rev.*, vol. 84, no. 1, pp. 98–107, 134, 2006.

[5] R. Hirt, N. Kühl, and G. Satzger, "An end-to-end process model for supervised machine learning classification: from problem to deployment in information systems," in *DESRIST 2017*, 2017.

[6] C. C. Aggarwal, T. J. Watson, R. Ctr, J. Han, J. Wang, and P. S. Yu, "A Framework for Clustering Evolving Data Streams," *Proc. - 29th int. conf. very large data bases*, pp. 81–92, 2003.

[7] A. Tsymbal, "The problem of concept drift: definitions and related work," Comput. Sci. Dep. Trinity Coll. Dublin, vol. 4, no. C, pp. 2004–15, 2004.

[8] S. Russell, D. Dewey, and M. Tegmark, "Research priorities for robust and beneficial artificial intelligence," Ai Mag., vol. 36, no. 4, pp. 105–114, 2015.

[9] J. Gama, I. Žliobaitė, A. Bifet, M. Pechenizkiy, and A. Bouchachia, "A survey on concept drift adaptation," ACM Comput. Surv., vol. 46, no. 4, pp. 1–37, 2014.

[10] I. Žliobaite, M. Pechenizkiy, and J. Gama, "An overview of concept drift applications," in Big Data Analysis: New Algorithms for a New Society, Springer, 2016, pp. 91–114.

[11] P. Kadlec and B. Gabrys, "Local learning-based adaptive soft sensor for catalyst activation prediction," AIChE J., vol. 57, no. 5, pp. 1288–1301, 2011.

[12] J. Han, M. Kamber, and J. Pei, Data Mining: Concepts and Techniques. 2012.

[13] M. I. Jordan and T. M. Mitchell, "Machine learning: Trends, perspectives, and prospects," Science. 2015.

[14] S. B. Kotsiantis, "Supervised Machine Learning: A Review of Classification Techniques," Informatica, 2007.

[15] H. Fromm, F. Habryn, and G. Satzger, "Service analytics: Leveraging data across enterprise boundaries for competitive advantage," in Globalization of Professional Services, 2012, pp. 139–149.

[16] K. Kambatla, G. Kollias, V. Kumar, and A. Grama, "Trends in big data analytics," J. Parallel Distrib. Comput., 2014.

[17] G. Widmer and M. Kubat, "Learning in the presence of concept drift and hidden contexts," Mach. Learn., vol. 23, no. 1, pp. 69–101, 1996.

[18] M. G. Kelly, D. J. Hand, and N. M. Adams, "The impact of changing populations on classifier performance," in ACM SIGKDD, 1999, pp. 367–371.

[19] I. Zliobaite, "Learning Under Concept Drift: An Overview," arXiv Prepr., 2010.

[20] G. I. Webb, R. Hyde, H. Cao, H. L. Nguyen, and F. Petitjean, "Characterizing concept drift," *Data Min. Knowl. Discov.*, vol. 30, no. 4, pp. 964–994, 2016.

[21] A. Ivannikov, M. Pechenizkiy, J. Bakker, T. Leino, M. Jegoroff, T. Kärkkäinen, and S. Äyrämö, "Online mass flow prediction in CFB boilers," in *Lecture Notes in Computer Science*, 2009, vol. 5633 LNAI, pp. 206–219.

[22] F. Fdez-Riverola, E. L. Iglesias, F. Díaz, J. R. Méndez, and J. M. Corchado, "Applying lazy learning algorithms to tackle concept drift in spam filtering," *Expert Syst. Appl.*, vol. 33, no. 1, pp. 36–48, 2007.

[23] S. J. Delany, P. Cunningham, and A. Tsymbal, "A comparison of ensemble and case-base maintenance techniques for handling concept drift in spam filtering," *LAIRS Conf.*, no. TCD-CS-2005-19, pp. 340–345, 2006.

[24] R. Giacomini and B. Rossi, "Detecting and predicting forecast breakdowns," *Rev. Econ. Stud.*, vol. 76, no. 2, pp. 669–705, 2009.

[25] M. Harries and K. Horn, "Detecting concept drift in financial time series prediction using symbolic machine learning," in *ACAI*, 1995.

[26] W. L. Perry, "Predictive policing: The role of crime forecasting in law enforcement operations." Rand Corporation, 2013.

[27] Y. Ding and X. Li, "Time weight collaborative filtering," in *CIKM '05*, 2005.

[28] M. Black and R. Hickey, "Detecting and adapting to concept drift in bioinformatics," in *Knowledge Exploration in Life Science Informatics*, Springer, 2004, pp. 161–168.

[29] R. Klinkenberg, "Meta-Learning, Model Selection, and Example Selection in Machine Learning Domains with Concept Drift," in *FGML*, 2005.

[30] A. Tsymbal, M. Pechenizkiy, P. Cunningham, and S. Puuronen, "Dynamic integration of classifiers for handling concept drift," *Inf. Fusion*, 2008.

[31] J. Zhou, L. Cheng, W. F. Bischof, and others, "Prediction and Change Detection in Sequential Data for Interactive Applications.," in *AAAI*, 2008, pp. 805–810.

[32] F. Fdez-Riverola, E. L. Iglesias, F. Díaz, J. R. Méndez, and J. M. Corchado, "Applying lazy learning algorithms to tackle concept drift in spam filtering," *Expert Syst. Appl.*, 2007.

[33] M. Black and R. Hickey, "Classification of customer call data in the presence of concept drift and noise," in *Lecture Notes in Computer Science*, 2002.

[34] G. Forman, "Incremental machine learning to reduce biochemistry lab costs in the search for drug discovery," in *BIOKDD*, 2002, pp. 33–36.

[35] B. Krawczyk, "Active and adaptive ensemble learning for online activity recognition from data streams," *Knowledge-Based Syst.*, 2017.

[36] S. Huang and Y. Dong, "An active learning system for mining time-changing data streams," *Intell. Data Anal.*, vol. 11, no. 4, pp. 401–419, 2007.

[37] S. Mohamad, A. Bouchachia, and M. Sayed-Mouchaweh, "A bi-criteria active learning algorithm for dynamic data streams," *IEEE Trans. neural networks Learn. Syst.*, 2016.

[38] I. Zliobaite, A. Bifet, B. Pfahringer, and G. Holmes, "Active learning with drifting streaming data," *IEEE Trans. Neural Networks Learn. Syst.*, vol. 25, no. 1, pp. 27–39, 2014.

[39] B. Kurlej and M. Wozniak, "Active learning approach to concept drift problem," *Log. J. IGPL*, 2011.

[40] B. Kurlej and M. Woźniak, "Learning curve in concept drift while using active learning paradigm," in *Adaptive and Intelligent Systems*, Springer, 2011, pp. 98–106.

[41] F. Mourão, L. Rocha, R. Araújo, T. Couto, M. Gonçalves, and W. Meira, "Understanding temporal aspects in document classification," in *WSDM '08*, 2008.

[42] M. Kukar, "Drifting concepts as hidden factors in clinical studies," in *Lecture Notes in Computer Science*, 2003.

[43] J. Ekanayake, J. Tappolet, H. C. Gall, and A. Bernstein, "Tracking concept drift of software projects using defect prediction quality," in *MSR'09.*, 2009, pp. 51–60.

[44] J. Luo, A. Pronobis, B. Caputo, and P. Jensfelt, "Incremental learning for place recognition in dynamic environments," in *IEEE/RSJ International Conference*, 2007, pp. 721–728.

[45] M. B. Harries, C. Sammut, and K. Horn, "Extracting hidden context," in *Machine Learning*, 1998.

[46] P. Gago, Á. Silva, and M. F. Santos, "Adaptive decision support for intensive care," in *Portuguese Conference on Artificial Intelligence*, 2007, pp. 415–425.

[47] A. Pawling, N. V Chawla, and G. Madey, "Anomaly detection in a mobile communication network," *Comput. Math. Organ. Theory*, vol. 13, no. 4, pp. 407–422, 2007.

[48] R. P. J. C. Bose, W. M. P. Van Der Aalst, I. Zliobaite, and M. Pechenizkiy, "Dealing with concept drifts in process mining," *IEEE Trans. Neural Networks Learn. Syst.*, 2014.

[49] L. Liao, D. J. Patterson, D. Fox, and H. Kautz, "Learning and inferring transportation routines," *Artif. Intell.*, vol. 171, no. 5–6, pp. 311–331, 2007.

[50] D. H. Widyantoro and J. Yen, "Relevant data expansion for learning concept drift from sparsely labeled data," *IEEE Trans. Knowl. Data Eng.*, vol. 17, no. 3, pp. 401–412, 2005.

[51] G. Lebanon and Y. Zhao, "Local likelihood modeling of temporal text streams," in *ICML*, 2008, pp. 552–559.

[52] Y. Xu, R. Xu, and W. Yan, "Power plant performance modeling with concept drift," in *IJCNN 2017*, 2017, pp. 2096–2103.

[53] Y. Sun, K. Tang, L. L. Minku, S. Wang, and X. Yao, "Online Ensemble Learning of Data Streams with Gradually Evolved Classes," *IEEE Trans. Knowl. Data Eng.*, 2016.

[54] K. Laghmari, C. Marsala, and M. Ramdani, "An adapted incremental graded multi-label classification model for recommendation systems," *Prog. Artif. Intell.*, 2018.

[55] V. Agrawal, B. Panigrahi, and P. M. V Subbarao, "Increasing Reliability of Fault Detection Systems for Industrial Applications," *IEEE Intell. Syst.*, 2018.

[56] S. G. Soares and R. Araújo, "An adaptive ensemble of on-line Extreme Learning Machines with variable forgetting factor for dynamic system prediction," *Neurocomputing*, 2016.

[57] M. Kubat, R. C. Holte, and S. Matwin, "Machine learning for the detection of oil spills in satellite radar images," *Mach. Learn.*, 1998.

[58] G. Ellis and F. Mansmann, "Mastering the Information Age Solving Problems with Visual Analytics." 2010.

[59] B. Krawczyk, L. L. Minku, J. Gama, J. Stefanowski, and M. Woźniak, "Ensemble learning for data stream analysis: A survey," *Inf. Fusion*, 2017.

[60] X. Zhu, Z. Ghahramani, and J. D. Lafferty, "Semi-Supervised Learning Using Gaussian Fields and Harmonic Functions," in *ICML*, 2003.

[61] J. Zerlang, "GDPR: A milestone in convergence for cyber-security and compliance," *Netw. Secur.*, vol. 2017, no. 6, pp. 8–11, 2017.