Proceedings of Machine Learning Research 105:1–19, 2019 Conformal and Probabilistic Prediction and Applications

Predicting with Confidence from Survival Data

Henrik Boström BOSTROMH@KTH.SE
School of Electrical Engineering and Computer Science, KTH Royal Institute of Technology, Sweden

Ulf Johansson ULF.JOHANSSON@JU.SE
Dept. of Computer Science and Informatics, Jönköping University, Sweden

Anders Vesterberg ANDERS.VESTERBERG@SCANIA.COM
Scania CV AB, Sweden

Editor: Alex Gammerman, Vladimir Vovk, Zhiyuan Luo and Evgueni Smirnov

Abstract

Survival modeling concerns predicting whether or not an event will occur before or on a given point in time. In a recent study, the conformal prediction framework was applied to this task, and so-called conformal random survival forest was proposed. It was empirically shown that the error level of this model indeed is very close to the provided confidence level, and also that the error for predicting each outcome, i.e., event or no-event, can be controlled separately by employing a Mondrian approach. The addressed task concerned making predictions for time points as provided by the underlying distribution. However, if one instead is interested in making predictions with respect to some specific time point, the guarantee of the conformal prediction framework no longer holds, as one is effectively considering a sample from another distribution than from which the calibration instances have been drawn. In this study, we propose a modification of the approach for specific time points, which transforms the problem into a binary classification task, thereby allowing the error level to be controlled. The latter is demonstrated by an empirical investigation using both a collection of publicly available datasets and two in-house datasets from a truck manufacturing company.

Keywords: Conformal prediction, survival modeling, random forests.

1. Introduction

Survival analysis studies the relationship between input features and the duration of time until an event of interest occurs, e.g., an adverse drug event or failure of a machine. One particular characteristic of survival data is that part of it may be *censored*, i.e., for some instances, e.g., patients or machines, the event is never observed. For such instances, we know that the event has not occurred until some point in time, but we have no further information beyond that point in time, e.g., for most patients the time of death is not (yet) known. Such observations are said to be *(right-)censored*. Particular care has to be taken to censored instances, since simply ignoring or removing them may introduce a significant bias, see e.g., (Hosmer and Lemeshow, 1999).

Many of the traditional approaches to survival modeling, such as Cox regression (Cox, 1972), make quite strong assumptions on the underlying distribution, and for this reason, non-parametric approaches have received increased attention. Examples of such methods are survival trees (Bou-Hamad et al., 2011) and random survival forests (Ishwaran et al., 2008). In particular the latter have become popular, mainly for the same reasons as (standard) random forests, including the effective handling of high-dimensional data and ease of parallelization. Additional reasons for their use are discussed in (Zhou and McArdle, 2015).

Similar to most standard machine learning algorithms, the output of random survival forests are, however, not necessarily well-calibrated, and the error levels cannot be guaranteed. Recently, an approach to applying random survival forests within the conformal prediction (CP) framework (Vovk et al., 2005), called *conformal random survival forests* (Boström et al., 2017), was presented. Similar to previous inductive conformal predictors, the approach relies on the assumption that calibration and test instances are drawn from the same, but unknown, underlying distribution. In contrast to previous approaches, each instance does here however not only consist of an input vector (x) and a (binary) label (y), but also a time point (t). The latter corresponds to either the time of the event (if $y = 1$) or censoring time (if $y = 0$). In (Boström et al., 2017), it was empirically shown that the error level of this model indeed is very close to the provided confidence level, and also that the error for predicting each outcome, i.e., event or no-event, can be controlled separately, by employing a Mondrian approach.

In most practical scenarios, one is however not interested in making predictions for the time points that are provided by the underlying distribution, but instead in making predictions with respect to some specific time point. For example, for training (and calibration), we may have been given a sample consisting of feature vectors representing trucks, together with their labels, e.g., failure or not, and the corresponding time of failure or censoring (depending on the label). When making predictions for test instances drawn from the same distribution, where feature vectors and time points are known, the probability that the set-valued predictions of the conformal survival forest excludes the correct label is bounded by the user-provided confidence level. However, if we are interested in making predictions for other time points than the ones provided by the underlying distribution, e.g., with respect to a fixed time point t_s, then the guarantee of the conformal prediction framework no longer holds. This follows from that we are now effectively considering a sample that is drawn from some other distribution than from which the calibration instances have been sampled. In this study, we propose a modification of conformal random survival forests for specific time points, and show that the error levels can still be controlled.

In the next section, we briefly describe conformal prediction and conformal random survival forests. In Section 3, we introduce an alternative approach which allows for considering specific time points, while maintaining the guarantee. In Section 4, we present results from applying the novel approach on a set of publicly available datasets and on two tasks concerning prediction of component failure using operational data from trucks. Finally, in Section 5, we discuss the main findings and outline directions for future work.

2. Preliminaries

2.1. Conformal prediction

Conformal prediction was originally developed for the transductive case (Gammerman et al., 1998), requiring re-training of the underlying model for each new instance to be predicted, something which often is computationally unfeasible. Inductive conformal prediction (ICP) was proposed as a computationally less costly approach (Papadopoulos et al., 2002), requiring only one underlying model to be generated, at the cost of having to set aside part of the training examples for calibration, which leaves less examples to use for model building. Below we briefly describe the ICP framework.

Let H be an underlying model and let $C = \{(x_1, y_1), \ldots, (x_n, y_n)\}$ be a calibration set, where each x_i is an input feature vector and $y_i \in Y$ is a class label, chosen from the set of possible labels Y. Let $A(H, x, y)$ be a *non-conformity measure*, i.e., a function returning a real-valued score for the infeasibility of y as a label for x according to H, with higher scores corresponding to higher infeasibility. Then a *p-value* for an instance x and label y is calculated with respect to the model H and the calibration set C in the following way:

$$p_{x,y} = \frac{l + (e+1)U[0,1]}{|C| + 1} \tag{1}$$

where $l = \sum_{(x_c,y_c)\in C} \mathbf{1}(A(H,x,y) < A(H,x_c,y_c))$ and $e = \sum_{(x_c,y_c)\in C} \mathbf{1}(A(H,x,y) = A(H,x_c,y_c))$, where $\mathbf{1}(\ldots)$ is the indicator function and $U[0,1]$ is a real number, randomly drawn from the uniform distribution $[0,1]$.

Given a confidence level c, the prediction $\boldsymbol{P}(x)$ for an instance x is the set of labels for which the corresponding p-values are equal to or higher than one minus the confidence level, i.e.,

$$\boldsymbol{P}(x) = \{y : y \in Y \;\&\; p_{x,y} \geq 1 - c\}$$

With no other assumptions than that the calibration and test instances are drawn from the same underlying distribution, independently from the training set and hence model generation, the probability of including the true class label in predictions for test instances is equal to or higher than the confidence level (Vovk et al., 2005; Vovk, 2012). It should be noted that this guarantee holds independently of the chosen non-conformity measure. To strengthen the guarantee to hold also for the individual labels, a class-conditional (or *Mondrian*) approach (Vovk et al., 2005) can be employed, where the calibration set is partitioned according to the class labels, and when calculating the p-value for a specific class label, only the corresponding partition of the calibration set is used.

2.2. Conformal random survival forests

2.2.1. The prediction task

When applying conformal prediction to the task of survival modeling, the most obvious difference to classification and regression is that each observation has a time stamp, i.e., the time at which an event occurred or the instance was censored. The approach proposed in (Boström et al., 2017) assumes that both training and test instances are drawn from some

specific, but unknown underlying distribution, where each instance is represented by an input vector, a time point, and a label (indicating whether an event or censoring occurred at the specified time point). The addressed task was then to predict what the correct label is for a randomly drawn test instance, when having access only to its input vector and time point.

2.2.2. The underlying model

A number of different split evaluation metrics have been proposed for survival trees and forests, such as the *log-rank* splitting rule (Ishwaran et al., 2008). Most of them (implicitly) assume a performance metric that evaluates each independent test instance over multiple points in time, e.g., all survival times that have been recorded, as done when employing e.g., Harrell's concordance index (C-index) (Ishwaran et al., 2008). The approach proposed in (Boström et al., 2017) instead focuses on a prediction task that considers one specific time point for each test instance, i.e., the survival or censoring time for that test instance. This allows for replacing previously proposed split evaluation metrics, which are computationally costly (due to considering multiple time points) with a less expensive split evaluation metric that only considers one time point per instance, namely a metric that minimizes the squared error of predicted survival probabilities in the two child nodes (assuming binary survival trees only). In addition, the approach requires each resulting child node to contain at least one event, and a node is turned into a leaf during tree growth if it contains no more than a specified number of training instances, or if none of the evaluated splits fulfill the previous condition.

2.2.3. Applying conformal prediction

For binary classification, a commonly employed non-conformity measure is to use the estimated probability that a label should not be assigned according to the underlying model, which in (Boström et al., 2017) was adapted to the survival modelling case:

$$A(H, x, t, y) = \mathbf{1}(y = 0)H(x, t) + \mathbf{1}(y = 1)(1 - H(x, t)) \ ,$$

where $H(x, t)$ is the estimated probability for that the event occurs ($y = 1$), conditioned on x, on or before time t. The definition of the p-value, also conditioned on time, i.e., $p_{x,t,y}$, is a straightforward adaptation of the definition of the standard p-value in Section 2.

Given the above non-conformity measure, together with an underlying model H (a random survival forest), a calibration set $C = \{(x_1, t_1, y_1), \ldots, (x_n, t_n, y_n)\}$, a confidence threshold c, the prediction $\boldsymbol{P}(x, t)$ for an input vector x and time point t (the test instance) is:

$$\boldsymbol{P}(x, t) = \{y : y \in \{0, 1\} \ \& \ p_{x,t,y} \geq 1 - c\}$$

2.2.4. Making predictions with respect to a specific time point

The practical limitation of not allowing the user to freely consider a time point of interest, in order to maintain the guarantee of conformal random survival forests, was already discussed in (Boström et al., 2017). This was further investigated in a master thesis (van Miltenburg,

2018) under the supervision of the two first authors of this paper. In the thesis, this limitation was empirically confirmed, showing that the error level may indeed exceed the specified level, if considering a specific time point for the test instances. Two modified versions of the original implementation of conformal random survival forests were considered in the thesis; one modifying the split metric to ignore instances censored prior to the time point of interest (while still considering them in the leaf nodes), and another heuristic approach, which sacrifices the guarantee on the error level for increased efficiency. In this work, we present an alternative, and novel, approach, which completely removes the need for generating survival trees, by resorting to (much faster) generation of binary classification trees, while still maintaining the guarantee of the conformal prediction framework.

3. Conformal Classification Forests for Survival Data

3.1. The prediction task

Similarly to (Boström et al., 2017), we assume that both training and test instances are drawn from some specific, but unknown underlying distribution, where each instance is a tuple (x, y, t), where x is an input vector, $y \in \{0, 1\}$ a label, and $t \in \mathbb{R}_{\geq 0}$ a time point, where the latter indicates the event time (if $y = 1$) or the censoring time (if $y = 0$). We are interested in correctly making predictions for test instances at a specified time point t_s, i.e., to predict whether an event has occurred or not up to (and including) time point t_s for each instance. It should be noted, however, that the correct label may not be known for all instances, namely whenever $t < t_s$ and $y = 0$ for an instance (x, y, t), i.e., the instance has been censored prior to the time point of interest. In the remaining cases, i.e., when $y = 1$ and $t \leq t_s$, or $t > t_s$, the correct labels can be derived, since in the first case, we know that the event has occurred on or before the time point of interest, while in the second case, we know that the event has not occurred before this time point. We modify the prediction task accordingly, and address the task of predicting the correct label for each test instance with a known label at a specified time point t_s. The problem may hence be viewed as a binary classification task, where predictions on instances with missing labels are ignored. The labels for this task are provided by the following function:

$$label((x, y, t), t_s) = \begin{cases} 1, & \text{if } y = 1 \text{ and } t \leq t_s \\ 0, & \text{if } t > t_s \\ missing, & \text{otherwise} \end{cases} \tag{2}$$

3.2. The underlying model

Since the prediction task ignores test instances with missing labels, a straightforward strategy, which is adopted here, is to also ignore such instances in the training set. It should be noted that some potentially valuable information may be lost by this strategy, and the investigation of alternative strategies that aim for exploiting also the unlabeled training instances is left for future research. The chosen strategy leads to that we end up with a set of binary labeled training instances, allowing us to choose freely among a large set of standard candidate algorithms to generate the underlying model, such as SVMs, ANNs, etc., rather than being limited to the less developed survival modeling algorithms. In this

work, we consider random forests (Breiman, 2001) as the underlying model, as they have frequently been shown to reach state-of-the-art performance, without requiring careful tuning of hyper-parameters. As a positive side-effect, they also allow for the conformal prediction framework to be applied without having to set aside a calibration set.

3.3. Applying conformal prediction

Since we are considering a random forest as the underlying model, a class probability distribution can straightforwardly be output for a test instance, by calculating the relative frequencies of the votes of the individual trees for the different class labels. We will use the predicted probability $H(x)$ that an event will occur ($y = 1$), given an input vector x, to obtain a non-conformity score:

$$A(H, x, y) = \mathbf{1}(y = 0)H(x) + \mathbf{1}(y = 1)(1 - H(x))$$

The use of out-of-bag predictions has been proposed as an alternative to using a separate calibration set when the underlying model is generated using bagging (Löfström et al., 2013; Boström et al., 2017), as done by random forests. This allows for using all available training data both for model construction and calibration. Similarly to (Boström et al., 2017), this approach is also adopted here.

4. Experiments

In this section, we present experimental results from applying conformal classification forests for survival data using first a collection of publicly available datasets and then two datasets from the automotive industry.

4.1. Experimental setup

In all experiments, we use forests with 500 binary classification trees as the underlying model, with 1/3 of the available features randomly selected for evaluation at each node during tree generation. We employ 10-fold cross validation and investigate the output of the resulting conformal predictors for the confidence level 0.95, for various specified time points. We consider both the standard approach of using a single calibration set and the class-conditional (Mondrian) approach, the latter with two separate calibration sets (one for each label), guaranteeing the error rate not only in general, but separately for each label.

4.2. Public datasets

In Table 1, the considered five publicly available datasets are listed, together with the number of instances, the number of features (not including time points and labels), fraction of instances for which an event has occurred, and the 25th percentile, the median and the 75th percentile for event and censoring times, respectively. As can be seen, all datasets are limited in size. It should be noted that close to 50% of the censored instances in the *grace* dataset are censored at time 180 (which is both the median and 75th percentile for the censoring times). Details of the datasets are described in (Hosmer and Lemeshow, 1999), except for the *pbc* dataset, which is described in (Kim et al., 2000).

Dataset	#instances	#features	Event frac.	Event time	Cens. time
actg320	1151	11	0.08	38/91/151	194/264/306
gbcs	686	8	0.25	601.5/868/1297.5	967/1499/1873.5
grace	1000	6	0.32	4/14.5/71	174/180/180
pbc	418	17	0.39	597/1083/2071	1419/2157/2863
whas	500	15	0.43	21/166/613	550/1245/1885

Table 1: Public datasets

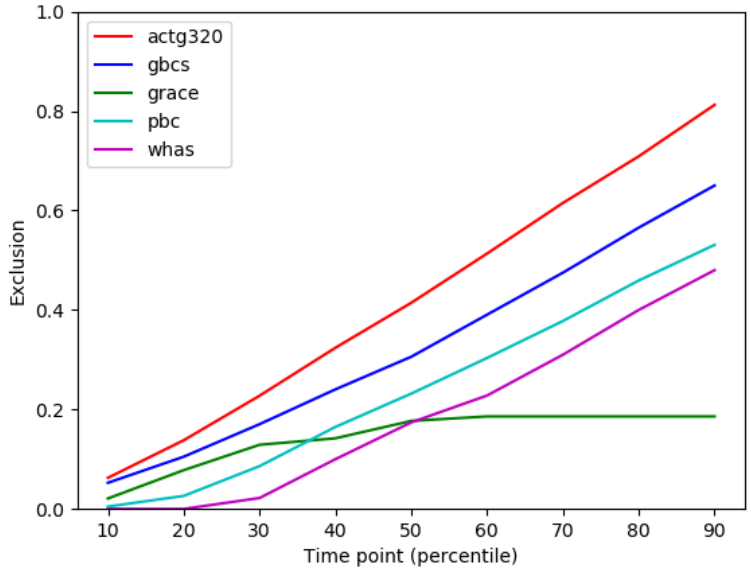

Figure 1: Proportion of excluded test instances over time

For each dataset, we consider nine fixed time points, corresponding to the 10th, ..., 90th percentile, respectively, of all (event or censoring) time points in the dataset. In Figure 1, the proportion of test instances not being labeled (*missing* according to Eq. 2) can be viewed for the (relative) time points and datasets (averaged over ten folds). Naturally, the proportion of excluded test instances increases with time, but levels off around the median for the *grace* dataset, as this corresponds to the latest observed time point in the dataset.

In Figure 2, the proportion of labeled test instances with a label corresponding to event is presented over time for all datasets. It can be seen that the class imbalance is initially reduced over time for all datasets, reaching a balance between the 60th and 80th percentile for three of the datasets, after which the imbalance starts to increase.

In Figure 3 and Figure 4, the accuracy and area under ROC curve (AUC) are displayed for all datasets over time. Accuracy, which in general benefits from class imbalance, clearly suffers from levelling out the class distributions. Note also that with the increased fraction of censored instances, less training data become available to the learning algorithm, which could explain why the accuracy does not eventually increase much. In contrast, the AUC,

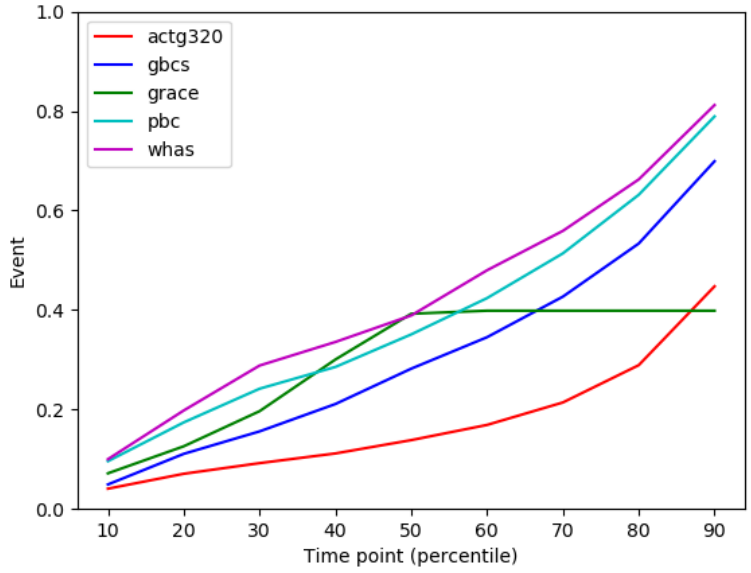

Figure 2: Proportion of included instances labeled with event

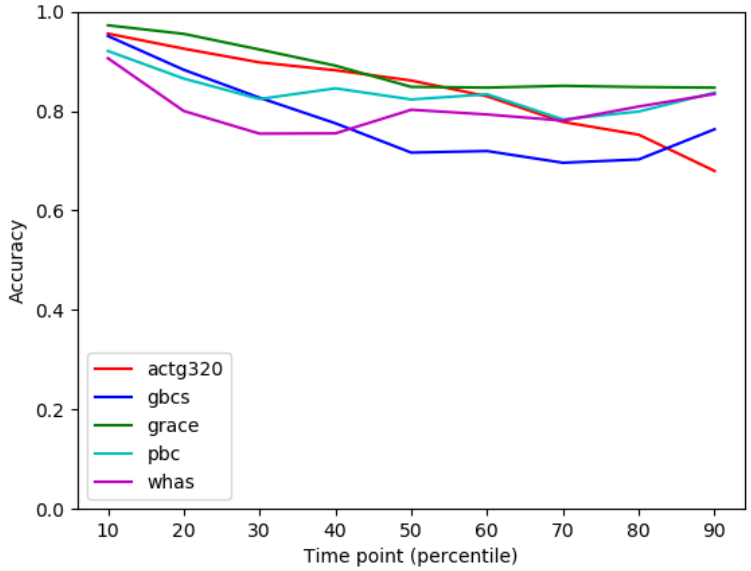

Figure 3: Accuracy over time

which in general is less affected by changes in the class distributions, is relatively stable over time.

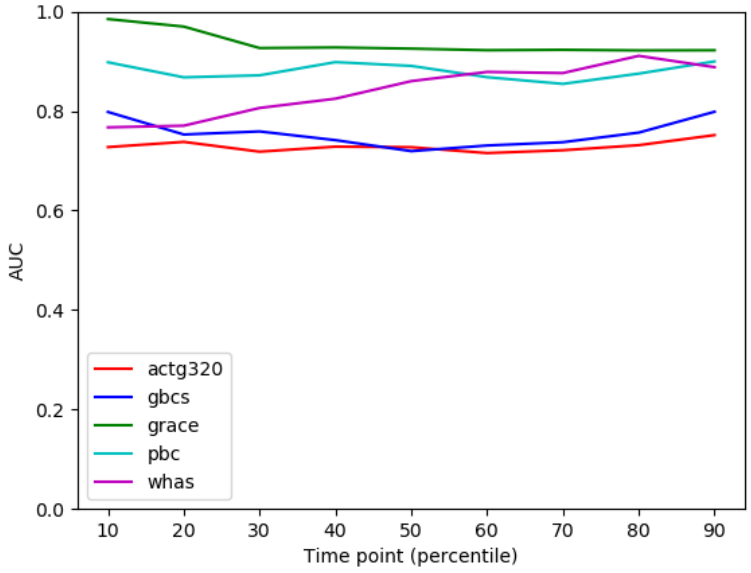

Figure 4: AUC over time

Now, turning the attention to the conformal predictors, in Figure 5 and Figure 6, the *validity*, i.e., fraction of test instances with known labels that are included in the prediction sets, and the *efficiency*, here measured by the number of labels in the prediction sets, are displayed for all datasets over time (again averaged over ten folds), using the confidence level 0.95 in all cases. One can see that the validity is indeed centered around the confidence level, although the variance increases with time, as a consequence of considering smaller sets of test instances with known labels (due to censoring). In Figure 6, a number of observations can be made. First, the standard approach, which does not provide a guarantee for the error level for each class separately, is generally more efficient than the class-conditional approach, which is in line with findings in previous studies. It can also be seen that for the standard approach, the informativeness of the predictions tend to decrease with time, which may be a reflection of the decreased accuracy (as seen in Figure 3). However, interestingly, this is less obvious for the class-conditional approach, for which there is no obvious pattern in the effect of time point on efficiency.

4.3. Case study I: NOx sensor failure prediction

4.3.1. Modeling task

Learning from the operational history of trucks to predict failure of components has been investigated quite extensively in the past, see e.g., (Frisk et al., 2014) for battery failure prediction, and (Prytz et al., 2015) for compressor failure prediction. Similarly to Boström et al. (2017), we here consider predicting failure of the NOx sensor, which provides a measurement of the nitrogen oxide concentration in the exhaust. The allowed levels of NOx

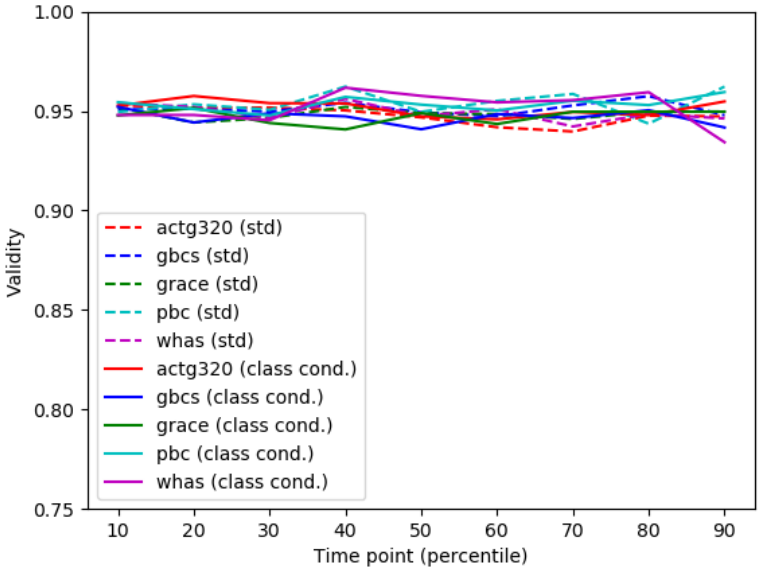

Figure 5: Validity over time

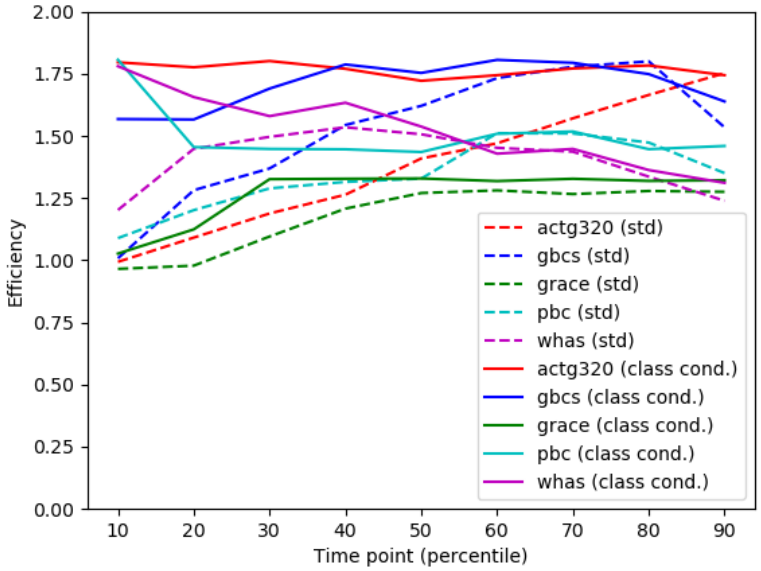

Figure 6: Efficiency over time

are strictly regulated[1], and a broken sensor means that the truck has to immediately visit a workshop, something which may not only result in costs associated with replacing the

1. http://www.eea.europa.eu/data-and-maps/indicators/eea-32-nitrogen-oxides-nox-emissions-1

(fragile and costly) component, but also in delivery delay, which often is associated with a penalty. Being able to accurately predict when the sensor is about to break, may help reduce these costs, by taking preventive actions, e.g., rescheduling maintenance and deliveries. However, unnecessarily taking a truck off the road is associated with other costs, due to sub-optimal use of resources. The errors hence need to be balanced properly.

In this study, we reuse the dataset studied in Boström et al. (2017). The dataset contains operational data records from a fleet of trucks manufactured for long haulage operations by Scania CV AB. The data has been collected on-board the trucks and extracted during visits at authorized workshops. Operational variables, such as fuel temperature, vehicle speed, ambient temperature, etc., are represented as histograms, with each bin value in a histogram denoting the frequency (under a given sample rate) that the truck has been operating within the conditions specified by the bin boundaries. For trucks with no observed NOx failure, a single snapshot (view of the data) is selected for each truck, which has been extracted at least seven days prior the latest snapshot. Trucks with a faulty NOx sensor are discovered from warranty claim records, and for these trucks, a snapshot is selected so that it occurs at least seven days before the day of breakdown. The time between the selected snapshot and the latest snapshot or breakdown is here taken as the censoring or event time, respectively. The final dataset includes 233 features, excluding the time points, which range from 7 to 1421 days. Out of the 16 980 instances (trucks) in the resulting dataset, 951 are labeled with an event (a problematic NOx sensor), which corresponds to a relative frequency of around 5%.

4.3.2. Experimental results

In Figure 7, the standard performance metrics accuracy (ACC) and area under ROC curve (AUC) are shown together with the fraction of censored instances in the test set (EXCL) and the proportion of test instances for which an event has occurred on or before each time point (EVENT). It can be seen how accuracy decreases with time, which is a natural consequence of the classification task becoming harder due to more uniformly distributed labels (as indicated by the growth of EVENT). AUC, on the other hand, stays relatively stable over time, similar to what was observed in the first experiment.

In Figure 8, the error levels are shown for test instances, on the various time points, They are quite consistent with the specified confidence level (95%), and they do not seem to be affected by the considered time points.

In Figure 9, it can be seen how the average number of elements in the output prediction sets vary with the considered time points. Consistent with previous studies (and above), the class-conditional approach again results in larger prediction sets. Interestingly, though, the efficiency is (slightly) improved over time, while the opposite holds for the standard approach. A similar pattern was observed also in the first experiment, and the reasons for this (still) remain to be found.

In Figure 10, it can be seen how the confidence relative to the predicted label vary with the considered time points for the two approaches (solid lines), and also how the confidence relative to the correct label (dashed lines) and incorrect labels (dotted lines) vary over time. For both the standard and the class-conditional approach, the confidence in the predictions that turned out to be correct, i.e., they agree with the test labels, is clearly higher than the confidence for predictions that turned out to be incorrect.

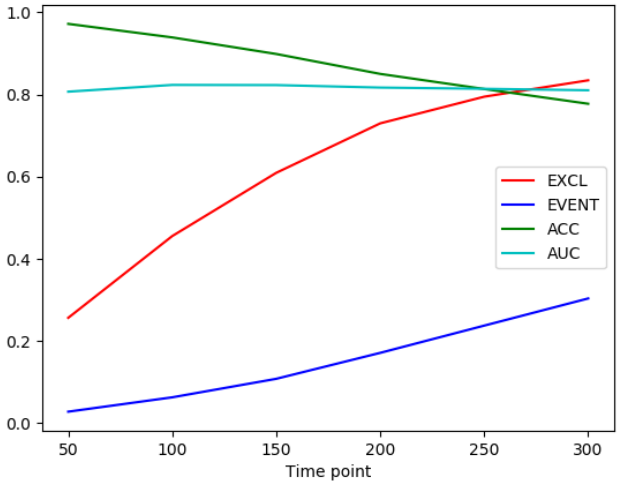

Figure 7: Some general statistics for NOx sensor failure prediction over time

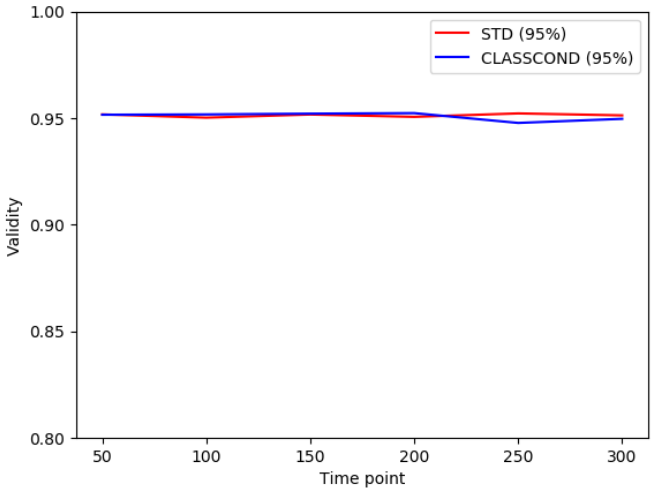

Figure 8: Validity for NOx sensor failure prediction over time

In Figure 11, it can be seen how the credibility in the predicted label vary with the considered time points for the two approaches (solid lines), and also how the credibility in the correct label (dashed lines) and incorrect labels (dotted lines) vary over time. Similarly to the results on confidence, the credibility of incorrect predictions are clearly lower than the credibility of correct predictions.

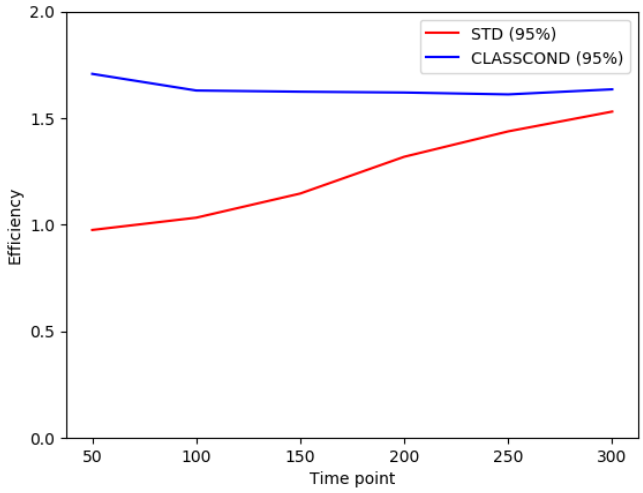

Figure 9: Efficiency for NOx sensor failure prediction over time

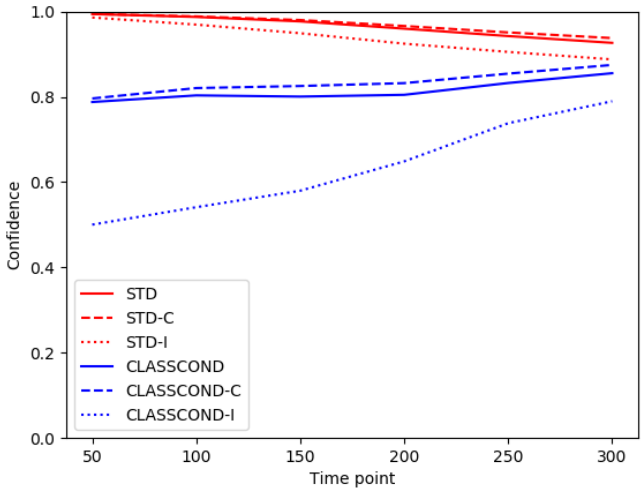

Figure 10: Confidence for NOx sensor failure prediction over time

4.4. Case study II: Alternator failure

4.4.1. Modeling task

An automotive charging system is made up of three major components: the battery, the voltage regulator and an alternator. The alternator works with the battery to generate power for the electrical components of a vehicle. An alternator wears out over time, but some factors like being under-specified, frequent starts and stops, high ambient temperature, moisture, salt, vibrations, and battery capacity affect the lifetime of the alternator.

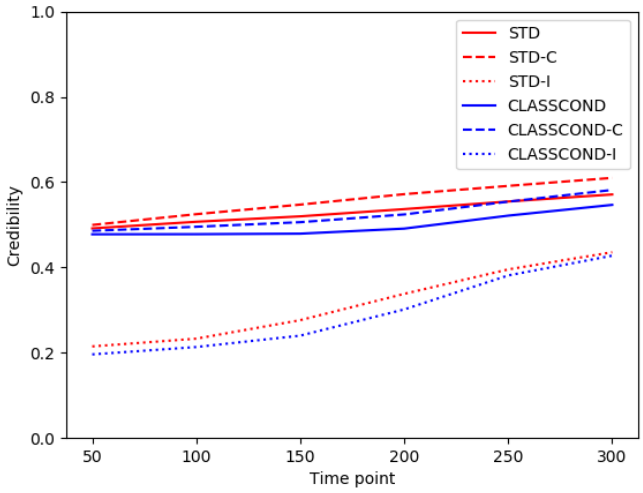

Figure 11: Credibility for NOx sensor failure prediction over time

Alternators are expensive to replace and therefore accurately predicting their failure is of great interest.

The dataset consists of 14 661 Scania buses and trucks, where the latter include distribution, construction and haulage trucks. 1158 out of the vehicles (7.9%) did have an alternator change within two years. The 45 variables represent operational data and selected specifications of how the vehicles were built. Included is also survival time, ranging from 1 to 377 weeks, and the event indicator. A snapshot containing operational data for a vehicle is selected so that it occurs at least seven days before the break down or censoring date.

4.4.2. Experimental results

Figure 12 shows the accuracy (ACC), the area under ROC curve (AUC), the fraction of censored instances in the test set (EXCL) and the fraction of events occurring on or before the time point of interest (EVENT). It can again be seen how accuracy initially decreases with time, but levels off (and even slightly increases) after a while, reflecting the class imbalance (as shown by EVENT). In contrast to the observations for the previous datasets, AUC is constantly increasing over time. A possible explanation for this could be that instances are not censored randomly, but that the cases that are more difficult to predict have a higher likelihood of being censored.

In Figure 13, the error levels are shown for test instances, on the various time points, As can be seen, the error levels are quite consistent with the specified confidence level, and they do not seem to be affected by the considered time points.

In Figure 14, it can be seen how the average number of elements in the output prediction sets vary with the considered time points. Consistent with previous studies (and above), the class-conditional approach again initially results in lower efficiency. However, for later time points this difference appears to diminish, and a pattern observed for this dataset is quite

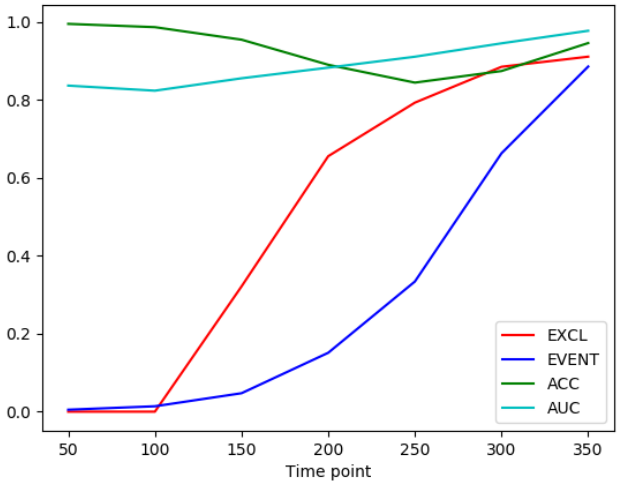

Figure 12: Some general statistics for alternator failure prediction over time

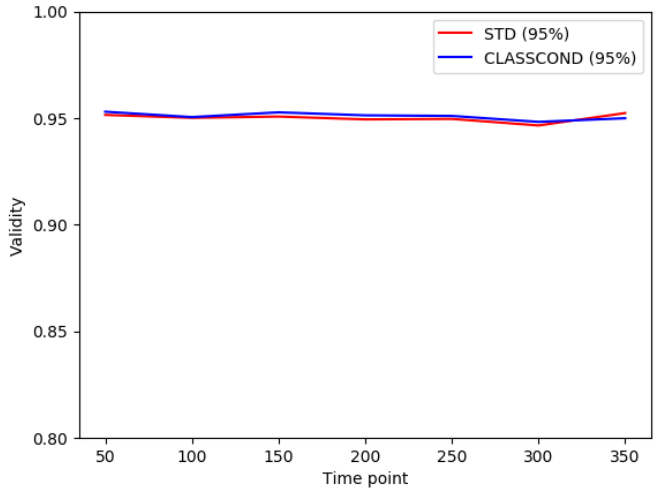

Figure 13: Validity for alternator failure prediction over time

different from the previous; the prediction sets increase in size for the standard approach only up to a point in time, after which the size decreases. In contrast, the prediction set size decreases constantly for the class-conditional approach.

In Figure 15 and 16, it can be seen how the confidence and credibility, respectively, vary with the considered time points for the standard and class-conditional approaches (solid lines), and also how they vary specifically for correct and incorrect predictions (dashed and dotted lines, respectively). It can be seen that the difference in confidence between correct

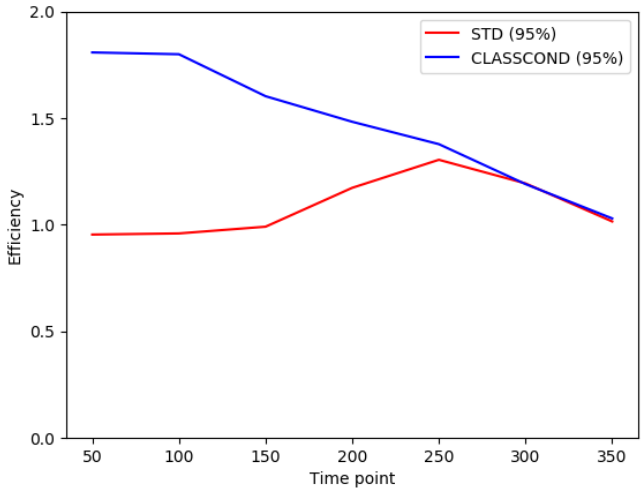

Figure 14: Efficiency for alternator failure prediction over time

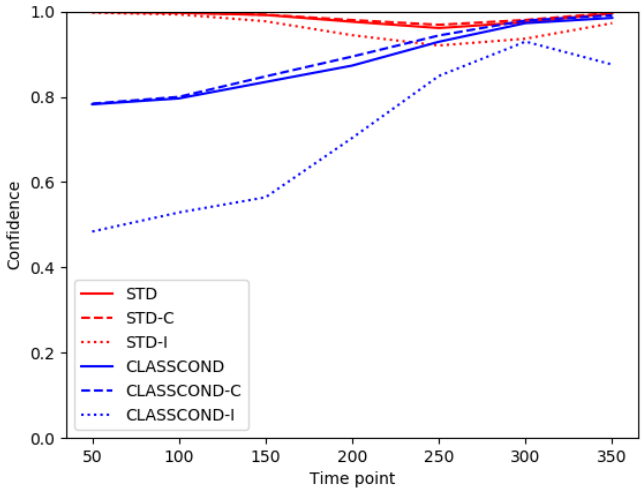

Figure 15: Confidence for alternator failure prediction over time

and incorrect predictions is more emphasized for the class-conditional approach, while the credibility is clearly much lower for the incorrect predictions for both approaches.

5. Discussion

In this study, we have presented an application of conformal prediction to the task of predicting whether an event has occurred or not on or before a specified point in time, for instances for which this is known, i.e., non-censored instances. We have presented

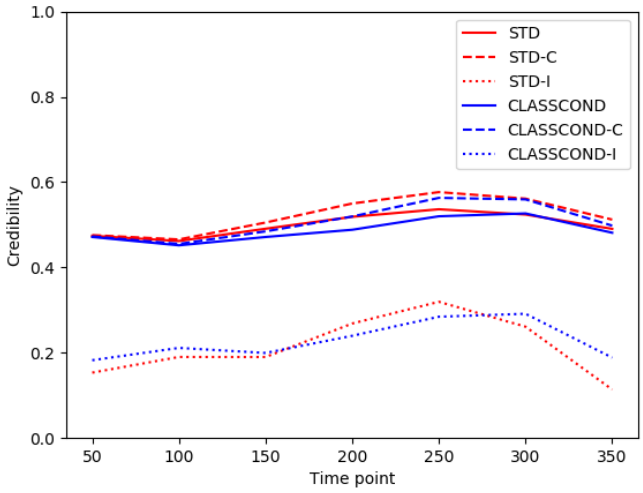

Figure 16: Credibility for alternator failure prediction over time

experimental results from using a set of publicly available datasets and two in-house datasets from the truck manufacturing company Scania. The results show that the difficulty of the classification task as measured by accuracy typically varies depending on the time point of interest. This can be explained by the effect of the chosen time point on the relative frequencies of the class labels, as well as on the amount of available (labeled) training data. The AUC, on the other hand, is relatively stable over time for all but one of the considered datasets, which may be explained by that AUC is resistant to changes in the class distribution. The observed improvement of AUC over time for one of the datasets indicates that censoring was not completely at random, but rather excluded cases that generally are more difficult to predict.

The results also show that the observed error levels are consistent with the employed confidence level, although a higher variance was observed for more distant time points of interest, again due to the smaller amounts of training (and test) data that were available at these time points. Similarly to what have been observed in previous studies, the class-conditional (Mondrian) approach tends to result in larger prediction sets, which seems to be an unavoidable price to pay for the class-specific guarantee. Quite surprisingly, and consistently for all considered datasets, the size of the prediction sets does however not seem to increase over time for the class-conditional approach, which contrasts to the standard approach, for which the size of the prediction sets tends to increase in general. Further work is needed to find an explanation for this observation.

It should be emphasized that the addressed task only considers instances for which it is known whether or not the event has occurred before the time point of interest. As a consequence, the guarantee from the conformal prediction framework is relevant only for instances that have not been censored prior to the time of interest, or in other words, instances for which it can be determined whether a label is correctly or incorrectly excluded. This may appear to be a reasonable restriction, and similar in spirit to some performance metrics

employed for evaluating survival models, such as Harrell's concordance index (C-index) (Ishwaran et al., 2008), which ignores comparisons that cannot be determined. However, when making a prediction for a specific instance, we will in general not know if it will be censored before the time point of interest, and hence we will not know if the guarantee applies to this specific instance. It is left for future research to investigate possible ways of extending the framework to provide guarantees also for censored instances. This could include using a separate model to predict (with confidence) whether or not the instance will be censored. Another possible approach is to instead use an additional label, e.g., *missing*, in the label set. These approaches will also, in different ways, allow for exploiting censored data during training, something which was pointed out above as a direction for future research.

Another direction for future research concerns applying also the Venn prediction framework to the task of predicting from survival data, in order to allow for outputting multi-probabilities, which are more directly useful in decision contexts, rather than p values. Such an application includes defining categories so that tight probability intervals may be obtained.

Acknowledgment

HB was supported by the Vinnova program for Strategic Vehicle Research and Innovation (FFI) Transport Efficiency.

References

Henrik Boström, Lars Asker, Ram Gurung, Isak Karlsson, Tony Lindgren, and Panagiotis Papapetrou. Conformal prediction using random survival forests. In *2017 16th IEEE International Conference on Machine Learning and Applications (ICMLA)*, pages 812–817. IEEE, 2017.

Henrik Boström, Henrik Linusson, Tuve Löfström, and Ulf Johansson. Accelerating difficulty estimation for conformal regression forests. *Ann. Math. Artif. Intell.*, 81(1-2): 125–144, 2017.

Imad Bou-Hamad, Denis Larocque, and Hatem Ben-Ameur. A review of survival trees. *Statist. Surv.*, 5:44–71, 2011. doi: 10.1214/09-SS047.

Leo Breiman. Random forests. *Machine Learning*, 45(1):5–32, 2001.

D. R. Cox. Regression Models and Life-Tables. *Journal of the Royal Statistical Society. Series B (Methodological)*, 34(2):187–220, 1972.

Erik Frisk, Mattias Krysander, and Emil Larsson. Data-driven lead-acid battery prognostics using random survival forests. In *Annual Conference of the Prognostics and Health Management Society 2014*, pages 92–101, 2014.

Alexander Gammerman, Volodya Vovk, and Vladimir Vapnik. Learning by transduction. In *Proceedings of the Fourteenth conference on Uncertainty in Artificial Intelligence*, pages 148–155. Morgan Kaufmann, 1998.

David W. Hosmer and Stanley Lemeshow. *Applied Survival Analysis: Regression Modeling of Time to Event Data*. John Wiley & Sons, Inc., New York, NY, USA, 1st edition, 1999.

H. Ishwaran, U.B. Kogalur, E.H. Blackstone, and M.S. Lauer. Random survival forests. *Ann. Appl. Statist.*, 2(3):841–860, 2008.

W Ray Kim, Russell H Wiesner, John J Poterucha, Terry M Therneau, Joanne T Benson, Ruud AF Krom, and E Rolland Dickson. Adaptation of the mayo primary biliary cirrhosis natural history model for application in liver transplant candidates. *Liver Transplantation*, 6(4):489–494, 2000.

Tuve Löfström, Ulf Johansson, and Henrik Boström. Effective utilization of data in inductive conformal prediction. In *Neural Networks (IJCNN), The 2013 International Joint Conference on*. IEEE, 2013.

Harris Papadopoulos, Kostas Proedrou, Volodya Vovk, and Alex Gammerman. Inductive confidence machines for regression. In *Machine Learning: ECML 2002*, pages 345–356. Springer, 2002.

Rune Prytz, Sawomir Nowaczyk, Thorsteinn Rgnvaldsson, and Stefan Byttner. Predicting the need for vehicle compressor repairs using maintenance records and logged vehicle data. *Engineering Applications of Artificial Intelligence*, 41:139 – 150, 2015.

Jelle van Miltenburg. Conformal survival predictions at a user-controlled time point: The introduction of time point specialized conformal random survival forests. Master's thesis, KTH Royal Institute of Technology, 2018.

Vladimir Vovk. Conditional validity of inductive conformal predictors. *Journal of Machine Learning Research - Proceedings Track*, 25:475–490, 2012.

Vladimir Vovk, Alex Gammerman, and Glenn Shafer. *Algorithmic Learning in a Random World*. Springer-Verlag New York, Inc., 2005.

Yan Zhou and John J. McArdle. Rationale and applications of survival tree and survival ensemble methods. *Psychometrika*, 80:811–833, 2015.