




PT-MESS: A Problem-Transformation Approach for Multi-Event Survival Analysis

Michela Venturini^{1,2}^a, Felipe Kenji Nakano^{1,2}^b and Celine Vens^{1,2}^c

¹*KU Leuven, Campus KULAK, Department of Public Health and Primary Care, Etienne Sabbelaan 53, 8500 Kortrijk, Belgium*

²*Itec, imec research group at KU Leuven, Etienne Sabbelaan 53, 8500 Kortrijk, Belgium*
{michela.venturini, felipekenji.nakano, celine.vens}@kuleuven.be


Keywords: Data Scarcity, Right-Censoring, Survival Analysis, Multi-Target Regression.


Abstract: Multi-event survival analysis is an under-explored field in literature, typically addressed by modeling each event independently or implying specific event settings. In this context, problem transformations approaches offer a promising alternative to rephrase the setting into standard multi-target regression. Nevertheless, they also suffer from the intrinsic presence of partial information in time-to-event data, since their application often requires the exclusion of censored observations, thus potentially discarding valuable information. In this work, we propose a novel Problem Transformation Approach for Multi-event Survival analysis (PT-MESS), which is capable of exploiting partial information, by encoding the survival outcome in a risk score based on the time-to-event distribution estimation. This approach allows the use of any multi-target machine learning model to address the original survival task. Using random forest as the underlying model, we conducted experiments using real-data multiple benchmarks from the medical domain and synthetic datasets. Our results revealed that PT-MESS provides superior or competitive results compared to competitors from the literature, especially when the events considered had a similar survival distribution.


1 INTRODUCTION

Survival analysis (SA) refers to a field of statistics that deals with time-to-event data, which often concerns medical applications where the outcome of interest is the time until occurrence of one (or more) adverse outcomes (e.g., death or cancer recurrence). SA is characterized by the presence of partial information, referred to as censoring, mainly due to individuals that are either lost to follow-up or do not experience the event during the follow-up, thus leaving the true time-to-event unknown. This phenomenon can be considered as a type of data scarcity as it leads to lack of labels and possibly poor prediction performance in case of high censoring rate. SA has been tackled either with algorithm-adaptation or problem-transformation machine learning approaches. While the first scenario implies adapting existing algorithms to handle survival data (Ping Wang, 2019), the latter one focuses on transforming time-to-event data to a well-studied problem, such as classification and re-

gression, enabling off-the-shelf models to be straightforwardly applied (Vock et al., 2016). For instance, Vock et al. (Vock et al., 2016) introduced an approach that employs weights to address SA as a binary classification task, however it excludes censored observations, potentially overlooking useful information. Despite the existence of numerous machine learning approaches for single-event SA, the literature presents few studies on the multi-event setting (Tjandra et al., 2021; Ishwaran et al., 2014). Furthermore, these studies have proposed algorithm-adaptation approaches which are solely tailored for competing and semi-competing risk analysis. That is, events are necessarily mutually exclusive (e.g., death from heart attack or breast cancer) or chronologically ordered (e.g., Alzheimer’s disease onset and death), thus being inadequate for more general settings. Although multiple outputs have been extensively studied in the machine learning literature (Xu et al., 2020), to date no problem-transformation approach exists to cast multi-event survival analysis to a multi-output prediction problem. In this work, we propose a problem-transformation approach, namely Problem Transformation Multi-Event Survival Analysis (PT-MESS),

^a <https://orcid.org/0000-0002-9947-0218>

^b <https://orcid.org/0000-0002-4884-9420>

^c <https://orcid.org/0000-0003-0983-256X>

which can be employed in any multi-event survival setting. More specifically, our approach relies on the survival distribution relative to each event to encode outcomes, allowing the inclusion of censored data in the model. Specifically, PT-MESS encodes time and censoring information into a score, enabling us to treat any multi-event SA task as a multi-target regression problem. By performing experiments on publicly available medical datasets, employing a standard multi-target regression model, we showcase that PT-MESS leads to superior or competitive results against methods from the literature in the majority of the cases.

2 PROPOSED METHOD

Survival outcome typically consists of two elements, namely time to event (or censoring) and binary status, indicating whether the individual experiences the event or is censored at the given time point. We propose to encode multi-event SA as a multi-target regression task, by transforming the survival outcome, for each event, into a single score. Specifically, the encoding process is based on the Kaplan-Meier estimate (Kaplan and Meier, 1958) of the survival curve relative to each event. Kaplan-Meier curves are used to describe the event distribution in the population, taking into account time and censoring information. Typically, multi-event survival problems are encoded as follows. Given N individuals and K events, each individual $i = 1, 2, \dots, N$ is associated with three variables: the covariates vector $\mathbf{x}^{(i)}$, the censoring statuses for each event $\mathbf{c}^{(i)}$ and the time information $\mathbf{o}^{(i)}$. In this context, the Kaplan-Meier curve estimate of the survival curve for each event k , $\hat{s}_k(t)$, is defined as follows:

$$\hat{s}_k(t) = \prod_{j: t_j \leq t} \left(1 - \frac{d_k^{(j)}}{r_k^{(j)}}\right) \quad (1)$$

with t_j a time when at least one event k happened, $d_k^{(j)}$ the number of events k that happened at time t_j , and $r_k^{(j)}$ the individuals still at risk for event k at time t_j , estimated from \mathbf{c}_k and \mathbf{o}_k . PT-MESS incorporates time and censoring information into the new outcome $m_k^{(i)}$ as follows:

$$m_k^{(i)} = 1 - \frac{\int_{t=0}^{o_k^{(i)}} \hat{s}_k}{\int_{t=0}^{\inf} \hat{s}_k} \quad (2)$$

where $\int_{t=0}^{o_k^{(i)}} \hat{s}_k$ is the restricted mean survival time, up to $o_k^{(i)}$, the time at which patient i experiences the

event or is censored, and $\int_{t=0}^{\inf} \hat{s}_k$ is the expected survival time of the population, for event k . Thus, Equation 2 can be seen as an indication of how early patient i is expected to experience event k w.r.t. the considered population. The higher the score, the lower the risk of experiencing the event. The new outcome for each individual i thus becomes $\mathbf{m}^{(i)}$, a vector containing a risk score for each event, that can be the outcome to any multi-target regression model. Given their efficiency and flexibility in handling high dimensional datasets, we employed multi-target random forest (Kocev et al., 2013) as the underlying regression model in our approach. Such model intrinsically exploits correlation among events in the splitting rule of the individual trees, by computing the average impurity reduction across them.

3 EXPERIMENTAL SETUP

3.1 Datasets

We employed 5 publicly available datasets, 3 of which are real world datasets from the medical domain (ADNI¹, MIMIC (Johnson et al., 2016) and CIBMTR²) and 2 were synthetic (scrData² and Synthetic). All of the real ones contain semi-competing risks, while the synthetic datasets contain both semi-competing risk and multi-event (events are not mutually exclusive and can happen in any order). Pre-processing of ADNI and MIMIC, and creation of Synthetic dataset were performed according to (Tjandra et al., 2021). Further details are reported at Table 1.

It can be seen that these datasets present different characteristics. Namely, MIMIC presents a considerable high number of features in comparison to ADNI. Similarly, CIBMTR contains a rather limited number of instances which are described by a considerably higher number of features. It can also be noticed that most of the datasets present scarce data where the censoring rate is frequently above 80%.

¹Data used in the preparation of this article were obtained from the Alzheimer’s Disease Neuroimaging Initiative (ADNI) database (adni.loni.usc.edu). The ADNI was launched in 2003 as a public-private partnership, led by Principal Investigator Michael W. Weiner, MD. The primary goal of ADNI has been to test whether serial magnetic resonance imaging (MRI), positron emission tomography (PET), other biological markers, and clinical and neuropsychological assessment can be combined to measure the progression of mild cognitive impairment (MCI) and early Alzheimer’s disease (AD).

²<https://cran.r-project.org/web/packages/SemiCompRisks/index.html>

Table 1: Characteristics of the datasets employed in this work where semi-competing datasets are represented using (S) and independent as (I). The terminal event in each dataset highlighted in bold.

Dataset (Setting)	Features	Instances	Events	Event (Censoring rate)
MIMIC (S)	13801	3822	3	1 (0.87), 2 (0.91) and 3 (0.92)
ADNI (S)	1917	1024	2	1 (0.82) and 2 (0.99)
CIBMTR (S)	9651	30	2	1 (0.82) and 2 (0.62)
ScrData (S)	2000	4	2	1 (0.34) and 2 (0.50)
Synthetic (I)	5000	15	2	1 (0.70) and 2 (0.71)

3.2 Comparison Methods

For a fair comparison, we focused on competitive methods that also employ ensembles of decision trees. To the best of our knowledge, the literature only presents a single study on random forests for multi-event SA (Ishwaran et al., 2014), nonetheless it is specifically tailored for competing risks, assuming that each patient can only experience one event. Hence, applying this method would require adaptations which are not straightforward. Further, the method proposed by Tjandra et al. was excluded (Tjandra et al., 2021). Although prominent, the authors have proposed a deep learning approach which requires time-consuming parameters tuning and lacks in interpretability. Thus, we compared the following methods:

- Random survival forest (**RSFi**): this approach learns a separate RSF (Ishwaran et al., 2008) per event and assumes that the time-to-events among all the events are independent.
- Inverse probability of censoring weighting (**IPCWi**): this problem transformation approach translates the survival task into a binary classification (Dong et al., 2020). Similarly to RSFi, this approach learns a separate model (random forest classifier) for each event.
- PT-MESS independent (**PT-MESSi**): A variant of our problem transformation approach, which learns a separate single-target random forest regressor for each event;
- (**PT-MESS**): A variant of our approach that builds a multi-target random forest regressor that considers all events at once;

Each ensemble model was trained with 200 trees and all other parameters were left to their default values³⁴. RSF is probably the most prominent approach using ensemble of trees, thus its comparison is mandatory. Similarly, IPCW is employed as comparison because it is a problem-transformation approach

comparable to ours, nonetheless it is worth mentioning that this method discards censored data. Moreover, both RSF and IPCW are originally designed for single-event applications. Thus, to make them comparable, it is necessary to build an independent model per event.

3.3 Evaluation

To estimate predictive performances of the models, we employed Harrell’s concordance index (Harrell et al., 1982) (C-index), one of the most used metrics in survival analysis. C-index is given by

$$C = \frac{\sum_{i,l} I(T_i > T_l) \times I(r_l > r_i) \times \Delta_l}{\sum_{i,l} I(T_i > T_l) \times \Delta_l} \quad (3)$$

where i and l refer to pairs of observations in the sample $i, l = 1, \dots, N$ with $i \neq l$, $r_i \in \mathbb{R}$ is the outcome risk score and T_i is the observed time-to event. Δ_l discards pairs of observations that are not comparable because the smaller survival time is censored. The C-index estimates how well a predicted risk score ranks observations according to their true time-to-event, taking into account censored data. Moreover, it is easy to interpret: $C = 0.5$ indicates a non-informative model prediction, while $C = 1$ indicates that the model is perfectly capable of separating patients with different outcomes. We evaluated our results based on per-event C-index as well as averaged C-index across events.

4 RESULTS

We present our results in Figure 1. All experiments were repeated using 5-fold cross-validation, stratified according to the rarest event (considering the censoring distribution) in the dataset. Average C-index (per single event, and overall) is reported, together with standard deviation.

As can be seen, our proposed method often has the upper-hand. More specifically, PT-MESS achieves the overall superior performance in 3 out of 5 datasets (CMBTR, scrData and Synthetic). Furthermore, when analyzing the performance per single event, our

³<https://scikit-learn.org/stable>

⁴<https://scikit-survival.readthedocs.io/en/stable>

proposed method surpassed the competitors in 7 out of 11 cases. As opposed to that, PT-MESSi managed, at its best, to be competitive with RSF and IPCW. We interpret this as an indication that PT-MESS is preferable over its counterpart and the competitors, considering both its performance and its computational complexity, as only one single model is required.

Additionally, we observed that PT-MESS outperformed its problem-transformation approach competitor, IPCWi, in several cases, as seen in events 1 and 2 of MIMIC, in event 1 of Synthetic and in the entire scrData dataset. As reported in Table 1, these datasets present a very high censoring rate. Hence, we may assume that PT-MESS is capable of correctly encoding the information of data which did not experience the event, whereas overlooking this data might lead to sub-optimal results.

Surprisingly, RSF and IPCWi yielded considerably better results in specific cases with very high censoring rate, as in MIMIC-event 3, and in the ADNI dataset. We believe that this is related to the low correlation of the events. As shown in Figure 2, their Kaplan Meier curves reveal that events in ADNI have very different survival curves (or time-to-event distributions). A similar behavior is observed in the MIMIC dataset where events 1 and 2 behave identically, whereas the curve associated to event 3 follows a significantly different tendency. This finding is further reinforced by the curve of scrData where the events appear to be substantially correlated, leading to superior results by our method.

5 CONCLUSIONS

We introduced a problem transformation approach to address multi-event survival analysis, where data scarcity is present as right-censored outcomes. Precisely, our approach encodes the typical survival outcome in a single score per event, based on Kaplan-Meier estimate of the survival curve and allows to include the censored observations in the model. As the underlying predictive model, we chose multi-target random survival forest. Our results revealed that PT-MESS is capable of providing superior results, indicating that predicting all events concurrently is beneficial over addressing them separately in the majority of the cases. In the challenging datasets, we could identify that our method, in its current form, struggles to predict events which are not correlated.

Hence, future work should aim to amend such deficiency. In this direction, we instigate the investigation of hybrid approaches, where global and local methods, may be used in cooperation to achieve su-

perior results. That is, we will extend our method to automatically detect and address correlated events during its building time, similarly to the concept of predictive bi-clustering trees used in multi-label classification (Zamith et al., 2020). Finally, we would also like to further validate our method by performing more experiments, specially regarding datasets with competing risks, as seen in (Ishwaran et al., 2014).

AUTHORS CONTRIBUTIONS

Conceptualization, M.V.; methodology, M.V. and F.K.N.; software, M.V.; writing—original draft preparation, M.V. and F.K.N.; writing—review and editing, C.V.; supervision, funding acquisition, C.V. All authors have read and agreed to the published version of the manuscript

FUNDING

This research was funded by the Research Fund Flanders (through research project G080118N and G0A2120N). The authors also acknowledge the Flemish Government (AI Research Program).

REFERENCES

- Dong, G., Mao, L., Huang, B., Gamalo-Siebers, M., Wang, J., Yu, G., and Hoaglin, D. C. (2020). The inverse-probability-of-censoring weighting (ipcw) adjusted win ratio statistic: an unbiased estimator in the presence of independent censoring. *Journal of biopharmaceutical statistics*, 30(5):882–899.
- Harrell, F., Califf, R., Pryor, D., Lee, K., and Rosati, R. (1982). Evaluating the yield of medical tests. *JAMA*, 247(18):2543–2546.
- Ishwaran, H., Gerds, T. A., Kogalur, U. B., Moore, R. D., Gange, S. J., and Lau, B. M. (2014). Random survival forests for competing risks. *Biostatistics*, 15(4):757–773.
- Ishwaran, H., Kogalur, U. B., Blackstone, E. H., and Lauer, M. S. (2008). Random survival forests. *The annals of applied statistics*, 2(3):841–860.
- Johnson, A. E., Pollard, T. J., Shen, L., Lehman, L.-w. H., Feng, M., Ghassemi, M., Moody, B., Szolovits, P., Anthony Celi, L., and Mark, R. G. (2016). MIMIC-III, a freely accessible critical care database. *Scientific Data*, 3(1):160035.
- Kaplan, E. L. and Meier, P. (1958). Nonparametric estimation from incomplete observations. *Journal of the American Statistical Association*, 53(282):457–481.

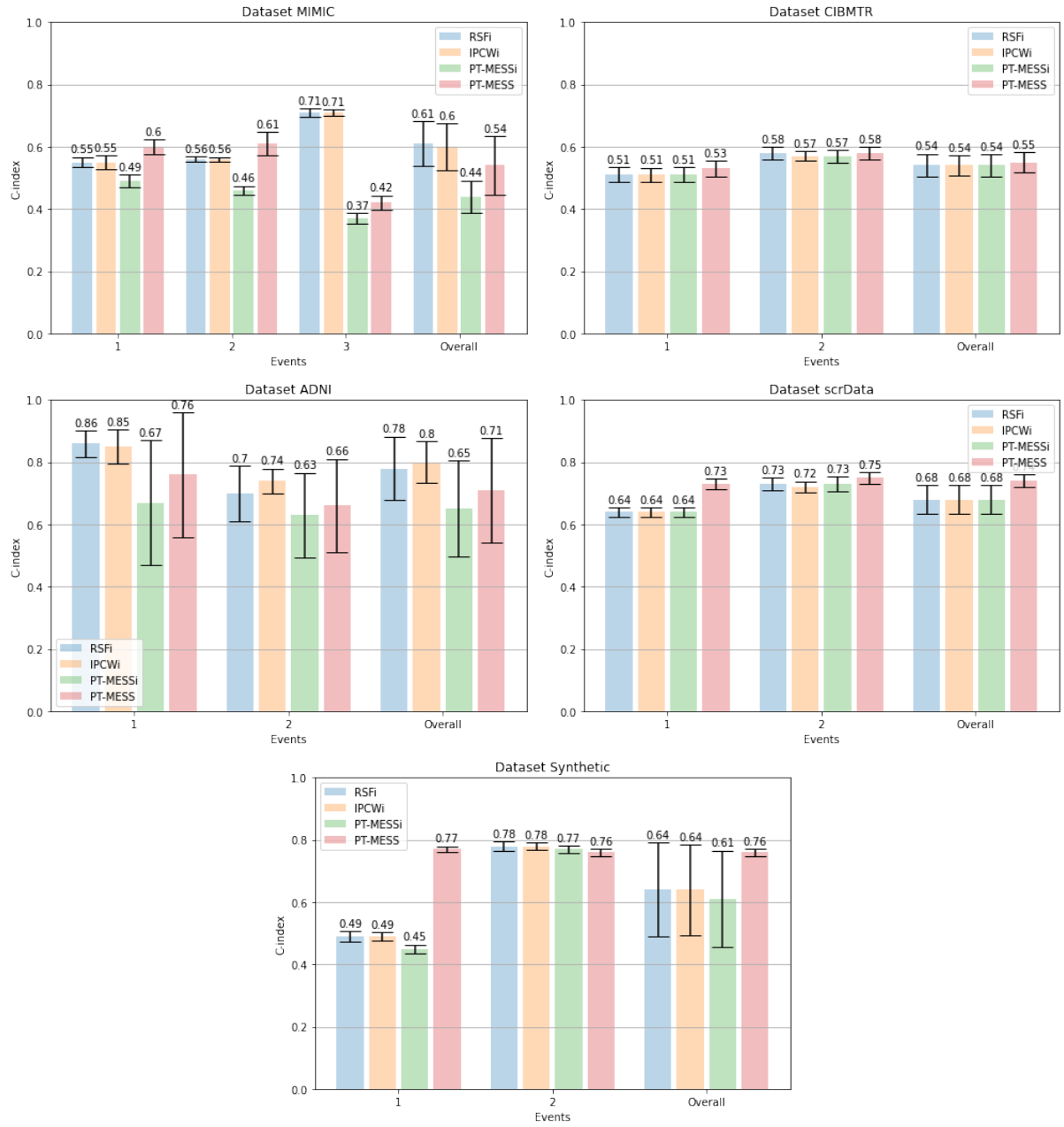


Figure 1: Results obtained on each dataset. We report the average performance obtained using the C-index and its standard deviation obtained using 5-fold cross validation.

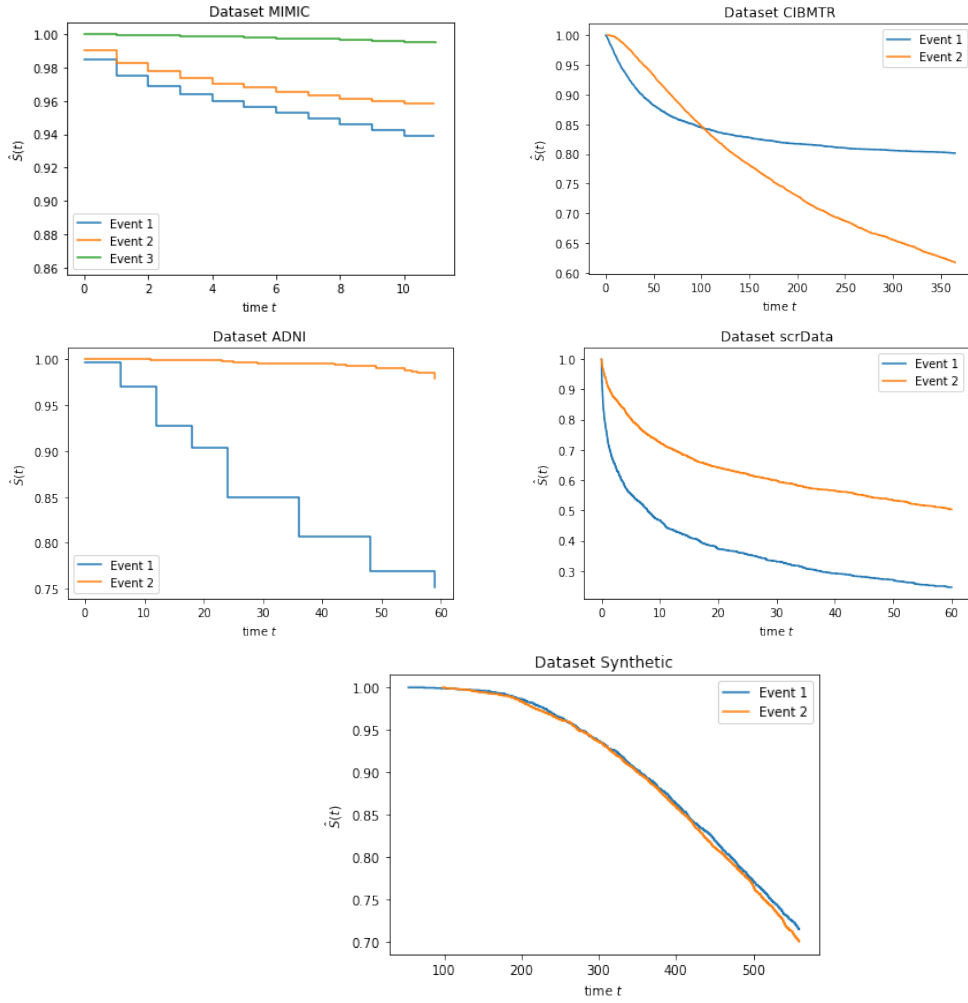


Figure 2: Kaplan-Meier estimate of the survival distribution for each event in the datasets: MIMIC, CIBMTR, ADNI and scrData, Synthetic.

Kocev, D., Vens, C., Struyf, J., and Džeroski, S. (2013). Tree ensembles for predicting structured outputs. *Pattern Recognition*, 46(3):817–833.

Ping Wang, Yan Li, C. K. R. (2019). Machine Learning for Survival Analysis: A Survey. *ACM Comput. Surv.*, 51(6).

Tjandra, D., He, Y., and Wiens, J. (2021). A hierarchical approach to multi-event survival analysis. *Proceedings of the AAAI Conference on Artificial Intelligence*, 35(1):591–599.

Vock, D. M., Wolfson, J., Bandyopadhyay, S., Adomavicius, G., Johnson, P. E., Vazquez-Benitez, G., and O’Connor, P. J. (2016). Adapting machine learning techniques to censored time-to-event health record data: A general-purpose approach using inverse probability of censoring weighting. *J Biomed Inform.*, 61:119–131.

Xu, D., Shi, Y., Tsang, I. W., Ong, Y.-S., Gong, C., and Shen, X. (2020). Survey on multi-output learning.

IEEE Transactions on Neural Networks and Learning Systems, 31(7):2409–2429.

Zamith, B., Nakano, F. K., Cerri, R., and Vens, C. (2020). Predictive bi-clustering trees for hierarchical multi-label classification. *ECML PKDD 2020*.