Predicting the length of a post-accident absence in construction with boosted decision trees

Anna Krawczyńska-Piechna[1*]

[1]Warsaw University of Technology, Faculty of Civil Engineering, Mechanics and Petrochemistry, 17 Łukasiewicza St., Płock, Poland

Abstract. Work safety control and analysis of accidents during construction performance are one of the most important issues of construction management. The paper focuses on post-accident absence as an element of occupational safety management. Somehow, the length of the post-accident absence can be treated as an indicator of building performance safety. The paper attempts to answer the question of whether it is possible to use boosted classifier ensembles to predict the post-accident absence length using a small set of historical observations, and which classification algorithm is the most promising to solve the prediction problem. It also proves that there is a dependence between the length of the post-accident absence and the cause of the accident or working conditions The choice of boosted algorithms is not accidental. Thanks to the use of aggregation methods it is possible to build classifiers that predict precisely and do not require any initial data treatment, which simplifies the prediction process significantly. The model of the prediction problem has been clarified. To identify the most promising classifier ensemble the prediction accuracy measures of selected classification algorithms were analyzed. The data used to build models was gathered on national (Polish) construction sites.

Keywords: classifier ensembles, post-accident absence, boosting

1 Introduction

An accident at work is defined in ESAW (European Statistics on Accidents at Work) methodology as a discrete occurrence during work which leads to physical or mental harm. Fatal accidents at work are those that lead to the death of the victim within one year, while non-fatal accidents at work collected within ESAW [1, 2] are those that imply at least four full calendar days of absence from work. For many years, in a majority of EU Member States, the highest incidence of accidents has been reported for persons employed in construction. In addition, for decades, more than half of all absence days registered are concentrated in only

* Corresponding author: Anna.Krawczynska@pw.edu.pl

© The Authors, published by EDP Sciences. This is an open access article distributed under the terms of the Creative Commons Attribution License 4.0 (http://creativecommons.org/licenses/by/4.0/).

three sectors of economic activity: manufacturing, construction and wholesale and retail trade. According to ILO [3] report, averagely, every 10 minutes one construction worker bears death during his work. This puts the construction industry at the head of the most hazardous professions basically due to the high variability of working conditions.

Accident statistics are being gathered and analyzed in all countries – in Poland by GUS (Main Statistical Office), in Europe by Eurostat, worldwide by ILO. However, collecting accidents' statistics itself is not the essence of the problem; it is identifying cause-and-effect relationships, the most common accidents' scenarios, factors affecting the severity of accidents, and occupational groups burdened with the highest risk of an accident. In Poland, such studies were conducted by Hoła's team, who developed computer knowledge database and a mathematical model of an accident event development in the construction industry [4, 5]. In turn, Drozd [6, 7] analyzed the costs of accidents and the length of a post-accident absence. A study on the above-mentioned papers and global research carried out by Saiful, Razwanul and Tarek [8], Mistikoglu et al. [9], Chua and Goh [10], or Chan, Leung and Liu [11] indicate that there is a strong need to build models of advisory systems supporting H&S management in its various aspects. Such systems should be effective, easy to use, and predict precisely with the historical data.

The paper matches up with work safety modelling problems. It attempts to answer the question whether it is possible to use boosted decision trees (a selected type of classifier ensembles) to predict the post-accident absence length using historical observations, and which algorithm is the most promising to apply in the prediction problem.

1.1 Why the post-accident absence and classifier ensembles?

The paper focuses on post-accident absence as an element of occupational safety management. According to Drozd [6], the length of the post-accident absence can be treated as an indicator of building performance safety. It is strongly affected by the working experience of the injured and the size of the company. In Polish conditions, the absence decreases with the increasing size of the company, which proves that larger construction companies pay more attention to works safety.

To solve the prediction problem classifier ensembles are suggested. In contrast to other prediction methods, classifier ensembles consist in building different models of the same phenomenon and then combining their judgments [12, 13]. The multi-model supervised learning in statistical data analysis has been used successfully in picture recognition, medical and biological sciences, customer segmentation, business modelling, etc. However, it is still less popular in project and construction planning, which is a pity, because aggregation algorithms, especially those based on decision trees or random forests, as Koronacki and Ćwik [14] claim, are thought to be the most effective classifiers. They do not require any initial data treatment, pre-processing, or analysis, which simplifies the prediction process significantly. They can also be used in case of missing data, what happens when collecting information on the construction site.

2 Mathematical model of the prediction problem

To predict the length of the post-accidence absence let's consider set U containing N observations which are different accidents at work. Each observation is described by a vector of attributes $[x_{i1}, x_{i2}, \ldots, x_{iL}, y_i]$. There are two kinds of attributes: predictors (called the input data) $X_1,\ldots, X_L$ and one target attribute Y (called the output data). Variables $x_{i1}, x_{i2}, \ldots, x_{iL}$, y_i describe attributes' values for the observation i ($i = 1,2,\ldots N$). The value of the target attribute is a class label. Therefore, set U can be defined as (1):

$$\left[x_i, y_i\right]_{N\times(L+1)} = \begin{bmatrix} X_1 & \cdots & X_L & Y \\ x_{11} & \cdots & x_{1L} & y_1 \\ x_{21} & \cdots & x_{2L} & y_2 \\ \cdots & \cdots & \cdots & \cdots \\ x_{N1} & \cdots & x_{NL} & y_N \end{bmatrix}. \quad (1)$$

Predictors are the circumstances of the accident. They are usually recorded to the accident report, and these are: the size of the company (the level of employment), work experience of the injured, type of work performed by the injured just before the accident, source of the accident, mechanization of works performed by the injured just before the accident. The target attribute is the length of the post-accident absence. The goal of the classification problem is to construct, using the historical data, a mathematical model that predicts the class of the unlabeled examples. This means that the dependence between the length of post-accident absence Y and accident's circumstances $\boldsymbol{X}=[X_1,\ldots, X_L]$ is being sought.

Classification can be done using a single classifier or a classifier ensemble, where a variety of classifiers (either different types of classifiers or different instantiations of the same classifier) are pooled before a final decision is made. Intuitively and mathematically, classifier ensembles provide an extra degree of freedom in the classical variance tradeoff, allowing solutions that would be difficult or impossible to reach with a single classifier [15]. There are several methods of combining classifiers: bagging (bootstrap aggregation), boosting, stacking, etc. In the present paper boosting technique is being analyzed. In this technique, learners are learned sequentially with early learners fitting simple models to the data. The training set used for each member of the series is chosen based on the performance of the earlier classifier in the series [16]. Observations wrongly classified by a single classifier receive a higher weight in order to be chosen to the next training set, so the algorithm is 'forced' to learn using them. The final classifier arises as a result of weighted component voting.

To build prediction models the data collected on 87 different construction sites in Poland and presented by Drozd [6] was used. The information about accidents was obtained from the statistical accident reports made after each accident. The observed predictors and the range of their historical values are collected in Table 1.

Table 1. Predictors and their historical values.

Predictor's name	Type of predictor's value	Observed predictors' values
company size (the level of employment)	numeric (number of company's employees)	from 4 to 241
experience of the injured workers	Numeric (number of years of work in construction)	from 1 to 16
type of work performed by the injured just before the accident	binary	1 in case of work at heights, 0 in other cases
the source of the accident	binary	1 – if the accident was caused by the worker himself (worker's psycho-physical state or improper behaviour, incl. not using protective equipment),

		0 – if the accident was caused by inappropriate workplace organisation and improper protection of the workplace
the state of mechanization of works performed by the injured just before the accident	binary	1 – if work was performed with equipment being moved or with a machine in motion, 0 – in other cases

In the examined data collection the target value (the length of the post-accident absence) varied between 2 and 29 days, while its mean value was 24.4 days and the standard deviation was 5.9 days. It should be mentioned here that, in Poland, if you are absent from work for more than 30 days, you are directed to repeat detailed medical examinations before returning to work. Therefore, for minor injuries, 29 or 30 days are usually the highest values. The problem that is being discussed and proposed methodology of its solution is of a discriminatory type. Therefore, the length of absence was divided into 6 separate classes (intervals): 0-5 days, 6-10 days, 11-15 days, 16-20 days, 21-25 days and 26-30 days.

3 The experiment

To answer the question of whether it is possible to use boosted decision trees to predict the post-accident absence length, and what is the prediction accuracy, two aggregation algorithms were examined: AdaBoost.M1 and Logit Boost. Five different classifiers were tested:

- Decision Stump – a weak, one-node decision tree,
- random tree classifier – a class for constructing a tree that considers randomly chosen attributes at each node and performs no pruning,
- J48 tree – a class for generating an unpruned or a pruned C4.5 decision tree,
- LMT tree – a classifier for building logistic model trees, which are classification trees with logistic regression functions at the leaves; the algorithm can deal with binary and multi-class target variables, numeric and nominal attributes and missing values;
- REPT tree – a fast decision tree learner, which builds a decision tree using information gain and prunes it using reduced-error pruning (with backfitting).

To obtain training sets it was acted in two ways: 70% dataset split and cross-validation with a different number of folds (incl. leave-one-out cross-validation) were performed. The maximum number of iterations was set to 100 and no resampling was allowed. All the calculations were performed in WEKA 3.8. environment.

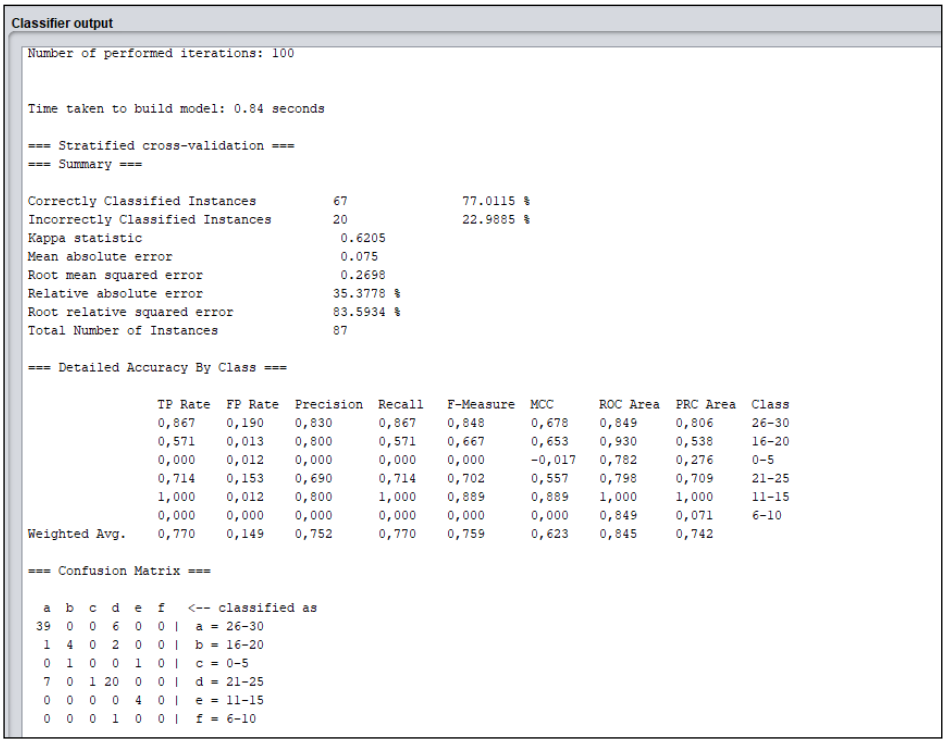
The best predicting algorithms are gathered in Table 2. The prediction accuracy is a percentage of correctly classified instances, while MAE stands for a mean absolute error.

Table 2. Comparison of the prediction accuracy obtained with different ensembles.

Classifier	**Max. prediction accuracy of a single classifier**	**Aggregation algorithm**	**Max. prediction accuracy of a classifier ensemble**	**MAE**
Decision Stump	58,6%	AdaBoost.M1	65,3%	0,23
		LogitBoost	77,0%	0,08
LMT	74,7%	AdaBoost.M1	75,8%	0,11
J48	72,4%	AdaBoost.M1	77,0%	0,08
Random tree	73,6%	LogitBoost	77,0%	0,08
REPT tree	69,0%	AdaBoost.M1	79,3%	0,12
		LogitBoost	79,3%	0,16

4 Discussion and conclusions

Accuracy measures for different ensembles and models built in Weka indicate that, whatever the loss function is, logit (LogiBoost algorithm) or exponential (AdaBoost.M1), classifier ensembles, as expected, provide better prediction accuracy than a single decision tree. The best improvement is observed for the weak Decision Stump classifier. Better prediction (at 81,6%) was obtained only with Random Committee aggregation algorithm and Random Tree classifier being aggregated, which was a result of supplementary calculations. Similar prediction accuracies, ca. 75-80%, were achieved within Random Forest method and for different bagging ensembles. It means that the value of ca. 80% accuracy is a ceiling value for the data being analyzed, whatever the aggregation method is. It's not bad but it's not excellent either. Moreover, when analyzing the confusion matrix and detailed accuracy measures, such as the F-measure and PRC Area, see Fig. 1, it was noted that the prediction accuracy for classes 0-5 and 6-10 is very low, regardless of the aggregation algorithm. Then, in the course of further research and in order to build a reliable decision model it is necessary to expand the knowledge database with observations for which the post-accident absence varies between 0 and 10 days.



Classifier output

```
Number of performed iterations: 100


Time taken to build model: 0.84 seconds

=== Stratified cross-validation ===
=== Summary ===

Correctly Classified Instances          67               77.0115 %
Incorrectly Classified Instances        20               22.9885 %
Kappa statistic                          0.6205
Mean absolute error                      0.075
Root mean squared error                  0.2698
Relative absolute error                 35.3778 %
Root relative squared error             83.5934 %
Total Number of Instances               87

=== Detailed Accuracy By Class ===

               TP Rate  FP Rate  Precision  Recall   F-Measure  MCC      ROC Area  PRC Area  Class
               0,867    0,190    0,830      0,867    0,848      0,678    0,849     0,806     26-30
               0,571    0,013    0,800      0,571    0,667      0,653    0,930     0,538     16-20
               0,000    0,012    0,000      0,000    0,000      -0,017   0,782     0,276     0-5
               0,714    0,153    0,690      0,714    0,702      0,557    0,798     0,709     21-25
               1,000    0,012    0,800      1,000    0,889      0,889    1,000     1,000     11-15
               0,000    0,000    0,000      0,000    0,000      0,000    0,849     0,071     6-10
Weighted Avg.  0,770    0,149    0,752      0,770    0,759      0,623    0,845     0,742

=== Confusion Matrix ===

  a  b  c  d  e  f   <-- classified as
 39  0  0  6  0  0 |  a = 26-30
  1  4  0  2  0  0 |  b = 16-20
  0  1  0  0  1  0 |  c = 0-5
  7  0  1 20  0  0 |  d = 21-25
  0  0  0  0  4  0 |  e = 11-15
  0  0  0  1  0  0 |  f = 6-10
```

Fig. 1. Classifier output for LogitBoost algorithm and Random Tree as a base classifier – an extract from Weka application window view.

The research proved that there is a strong dependence between the length of the post-accident absence and selected accident's circumstances. It also proved that classifier ensembles based on boosting algorithms are a valuable tool to support the prediction of the length of the post-accident absence, even if the set of historical data used to learn the classifier is relatively small or there is data missing. The most time-effective algorithms: boosted J48 decision tree and Decision Stump will be analyzed in further research.

Reference

1. European Statistics on Accident at Work (ESAW), *Summary methodology. Eurostat Methodologies and Working Papers* (Publications Office of the European Union, Luxembourg, 2013)
2. European Statistics on Accident at Work (ESAW), Accidents at work statistics (2016), http://ec.europa.eu/eurostat/statistics-explained/index.php?title=Accidents_at_work_statistics
3. International Labor Organization (ILO), A Vision for Sustainable Prevention, XX World Congress on Safety and Health at Work, Global Forum for Prevention, 24-27 August 2014, Frankfurt, Germany (2014)
4. B. Hoła, M. Szóstak, Analysis of the State of the Accident Rate in the Construction Industry in European Union Countries, Arch. Civ. Eng. **61(4)**, 13-34 (2015)
5. M. Szóstak, *Modelling of the development of an accident situation in the construction industry*, PhD Thesis (Wroclaw University of Technology, Wroclaw, 2017)
6. W. Drozd, *Regression analysis of accident absenteeism and variables describing working conditions* (Cracow University of Technology Publishing, Cracow, 2015)
7. W. Drozd, Analysis of Cost Regression and Post-Accident Absence, AIP Conference Proceedings **1863(1)**, 230004 (2017)
8. M. Saiful, I. Razwanul, M. Tarek, Safety Practices and Causes of Fatality in Building Construction Projects: A Case Study for Bangladesh, Jordan J. Civ. Eng. **11 (2)**, 267-278 (2017)
9. G. Mistikoglu, I.H. Gerek, E. Erdis, P.E. Mumtaz Usmen, H. Cakan, E.E. Kazan, Decision tree analysis of construction fall accidents involving roofers, Expert Syst. Appl. **42(4)**, 2256-2263 (2015)
10. D.H.K. Chua, Y.M. Goh, Incident Causation Model for Improving Feedback of Safety Knowledge, J. Constr. Eng. Manag. **130(4)**, 542-551 (2004)
11. I.Y.S. Chan, M.Y. Leung, A.M.M. Liu, Occupational health management system: A study of expatriate construction professionals, Accid. Anal. Prev. **93**, 280-290 (2016)
12. M. Walesiak, E. Gatnar, *Data analysis in R* (in Polish) (PWN, Warsaw, 2009)
13. E. Gatnar, *Multi-model approach in discrimination and regression* (in Polish) (PWN, Warsaw, 2009)
14. J. Koronacki, J. Ćwik, *Statistical learning systems* (in Polish) (Akademicka Oficyna Wydawnicza Exit, Warsaw, 2008)
15. N.C. Oza, K. Tumer, Classifier ensembles: Select real-world applications, Inform. Fusion, **9(1)**, 4-20 (2008)
16. D. Opitz, R. Maclin, Popular Ensemble Methods: An Empirical Study, Int. J. Artif. Intell. Res. **11(2)**, 267-278 (1999)