Automation in Construction 69 (2016) 102–114

Contents lists available at ScienceDirect

Automation in Construction

journal homepage: www.elsevier.com/locate/autcon

Application of machine learning to construction injury prediction

Antoine J.-P. Tixier [a,*], Matthew R. Hallowell [a], Balaji Rajagopalan [b], Dean Bowman [c]

[a] Department of Civil, Environmental, and Architectural Engineering, University of Colorado at Boulder, Boulder, CO 80309, United States
[b] Department of Civil, Environmental, and Architectural Engineering, Cooperative Institute for Research in Environmental Sciences (CIRES), University of Colorado at Boulder, Boulder, CO 80309, United States
[c] Bentley Systems, United States

ARTICLE INFO

Article history:
Received 10 October 2015
Received in revised form 24 March 2016
Accepted 22 May 2016
Available online 15 June 2016

Keywords:
Machine learning
Construction safety
Predictive modeling
Injury prevention
Random Forest
Boosting
Attribute

ABSTRACT

The needs to ground construction safety-related decisions under uncertainty on knowledge extracted from objective, empirical data are pressing. Although construction research has considered machine learning (ML) for more than two decades, it had yet to be applied to safety concerns. We applied two state-of-the-art ML models, Random Forest (RF) and Stochastic Gradient Tree Boosting (SGTB), to a data set of carefully featured attributes and categorical safety outcomes, extracted from a large pool of textual construction injury reports via a highly accurate Natural Language Processing (NLP) tool developed by past research. The models can predict *injury type*, *energy type*, and *body part* with high skill ($0.236 < RPSS < 0.436$), outperforming the parametric models found in the literature. The high predictive skill reached suggests that injuries do not occur at random, and that therefore construction safety should be studied empirically and quantitatively rather than strictly being approached through the analysis of subjective data, expert opinion, and with a regulatory and managerial perspective. This opens the gate to a new research field, where construction safety is considered an empirically grounded quantitative science. Finally, the absence of predictive skill for the output variable *injury severity* suggests that unlike other safety outcomes, *injury severity* is mainly random, or that extra layers of predictive information should be used in making predictions, like the energy level in the environment. In the context of construction safety analysis, this study makes important strides in that the results provide reliable probabilistic forecasts of likely outcomes should an accident occur, and show great potential for integration with building information modeling and work packaging due to the binary and physical nature of the input variables. Such data-driven predictions had been absent from the field since its inception.

1. Introduction and motivation

Construction is one of the largest industries in the United States, but is also one of the deadliest [12]. Between 1992 and 2010, an average of 730 lives have been claimed each year [20]. Despite the numerous efforts that have been motivated by this alarmingly poor performance, injury statistics have not significantly improved in the past decade [12]. This might be explained by the fact that the construction industry has reached saturation with respect to traditional approaches to safety and that innovations are needed [27]. Risk analysis has emerged as a promising alternative to managerial and regulation-based approaches. However, construction safety risk analyses are currently limited because existing techniques overlook the complex and dynamic nature of construction sites and are not based on empirical data.

To jointly address these limitations, Esmaeili and Hallowell [26,28] laid the groundwork of a new conceptual framework, offering a systematic and comprehensive way to extract safety critical structured information from unstructured injury reports. Unlike traditional safety risk analysis techniques, this attribute-based approach renders construction injuries as the resulting outcome of the joint presence of a worker and the interplay among a finite set of universal descriptors of the work environment that are observable before an injury occurs. These binary attributes, also called injury precursors, make physical sense and are related to construction means and methods, human behavior, and environmental conditions. For instance, in the following excerpt of an injury report: "employee was welding and grinding inside tank and experienced discomfort to left eye", four fundamental attributes can be identified: (1) *welding*, (2) *grinding*, (3) *tank*, and (4) *confined workspace*.

The attribute-based framework derives its strength from the ability to capture and encode the information of every possible construction situation in a finite, standardized format, regardless of trade, project type, or industry sector. Therefore, as illustrated in Fig. 1, extracting attributes and various safety outcomes from injury reports (i.e., objective empirical data) enables the constitution of a structured, consistent

* Corresponding author.
E-mail addresses: antoine.tixier-1@colorado.edu (A.J.-P. Tixier), matthew.hallowell@colorado.edu (M.R. Hallowell), rajagopalan.balaji@colorado.edu (B. Rajagopalan), dean.bowman@bentley.com (D. Bowman).

http://dx.doi.org/10.1016/j.autcon.2016.05.016
0926-5805/

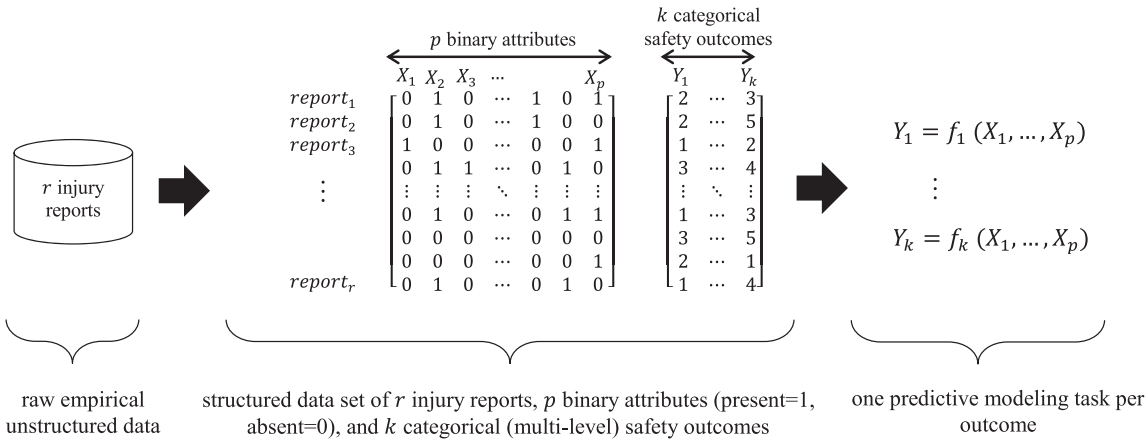

Fig. 1. The derivation of predictive models from injury reports is enabled by the attribute-based framework.

multivariate data set ideally suited for data mining, predictive modeling, and, thus, for knowledge discovery. Such new knowledge can enhance understanding of the underlying mechanisms that shape construction safety risk and create injuries. More precisely, *this study seeks to demonstrate that the workflow illustrated in Fig. 1 is viable and can be used to produce empirically-driven models with high predictive skill.* A fundamental postulate made here is that construction safety is not a strictly managerial outcome, but rather features a non-random component that can be studied by means of observation, like any other natural phenomenon. If this assumption holds, adopting the attribute-based framework would succeed in transforming construction safety research from opinion-based and qualitative to objective, empirically grounded quantitative science.

The effectiveness of the attribute-based framework depends on a number of methodological parameters including: (1) the way attributes are created and defined, (2) the quality and quantity of the injury reports available, (3) the technique with which attributes are extracted from the reports, and (4) the methods used for data mining and predictive modeling. As will be discussed in the background section, all previous work in this emerging research area (e.g., [22,26,28,29,30,70]) is subject to limitations with respect to one or more of the aforementioned parameters.

Building on three recent studies [22,66,70] that respectively addressed the limitations pertaining to the first three of the aforementioned criteria, here we tackle the limitations related to the fourth: predictive modeling. More specifically, two state-of-the-art machine learning (ML) algorithms, Random Forest (RF) and Stochastic Gradient Tree Boosting (SGTB), were used to predict safety outcomes from fundamental construction attributes. As will be shown, the models built outperform that of past research, in terms of predictive skill, variety of outcomes predicted, and actionable feedback that can be used to direct efforts towards targeted preventive actions and corrective measures.

2. Background and point of departure

This section provides the inspiration for our work, a brief description of past research in the domain of attribute-based safety analysis and in the application of ML in the construction industry, and the expected contributions.

2.1. Why does prediction of safety outcome matter?

Many industries, including construction, struggle with decision-making under uncertainty. Making the wrong decisions can have dramatic consequences, especially when lives are at stake. In healthcare, for example, Seera and Lim [58] observed that lack of experience, information overload, and unawareness of the most recent advancements in medical research were the leading causes of misdiagnosis by physicians. In the exact same way, even an experienced construction worker or safety manager has limited personal history with accidents. They may have witnessed, in their entire professional life, hundreds of near misses and first aid injuries, dozens of medical cases and lost work time injuries, and, perhaps, a few permanent disablement injuries and fatalities. Because of this limited experience with incidents, they may misdiagnose the risk of a given construction situation. It is actually well known that poor hazard recognition skill is a proximal cause of risk misperception and injury in construction [2,13]. People working upstream of the construction phase, like designers, face an even greater risk of failing to recognize hazards and misestimating risk[2,8].

Furthermore, without even considering the limited experience problem, human judgment and intuition will always be subject to important biases and fallacies (e.g., [69]). Also, humans have very limited capability of inducing knowledge from large numbers of observations [59]. This is due to the fact that human short-term memory is only capable of handling at most seven items evaluated for seven attributes at the same time [50].

On the other hand, ML can induce general rules from very large amounts of cases belonging to highly dimensional spaces, and is therefore a way to ground safety-related decisions under uncertainty on empirical knowledge. This could lead to improved decision-making and save lives. Indeed, other industries have begun to realize great benefits by transitioning from subjective to objective decision-making thanks to statistical learning. For instance, Seera ad Lim [58] trained ML models on large numbers of health records to automatically diagnose new patients, providing physicians with an opportunity to reconsider initial decisions and improve diagnosis accuracy.

2.2. Limitations of previous work on attribute-based construction safety

Although Esmaeili and Hallowell [26,28] made important strides by introducing and using the attribute-based framework for the first time, some serious limitations remained. In particular, some of the attributes identified via manual content analysis were not in full accordance with the framework as they were outcomes (e.g., *structure collapse*, *falling from roof*). By nature, an injury precursor should be observable *before* an injury occurs. Some other attributes were overlapping (e.g., *working underground*, *working in a confined space*), or loosely defined (e.g., *not considering safety during site layout*). Finally, the content analysis had rather low consistency (76% of inter-coder agreement), and only 300 reports all related to high severity struck-by injuries were analyzed, so only part of the picture was captured.

Esmaeili et al. [29] took the research a step further by using commercial software to automatically extract attributes from a larger amount of reports (1450). However, the low accuracy of the procedure (21%