25-28 June 2016 Hotel Danubius Health Spa Resort Margitsziget****, Budapest, Hungary

Creative Construction Conference 2016

Severity Prediction Models of Falling Risk for Workers at Height

Hainan Chen, Xiaowei Luo*

Department of Architecture and Civil Engineering, City University of Hong Kong, Hong Kong SAR

Abstract

Construction industry has one of the highest accident and fatality rates among other major industries, with more than 60,000 fatal accidents each year worldwide. Falling from height is one of the leading causes of fatalities and injuries in construction. Passive protection devices (e.g., safety net) have been used to minimize the impact of falling from height for ages, while proactive warning systems appear recently to alert the workers when they are at risks of falling. To provide appropriate warnings to the worker but not to distract them due to the false alarm, the falling risk needs to be carefully evaluated. In this paper, the authors introduced algorithms for falling risk prediction and evaluated their performance. Injuries records during 2005 to 2015 were extracted from the OSHA database and 1161 intact falling-related record were used in this study. K-Modes, RBF network and Decision Trees are chosen to build three risk prediction models, and the performance of those three proposed models were evaluated using the OSHA injuries record data. The results indicate that the DT-based falling risk prediction model has the best performance of 75% and the top three critical factors of falling event's severity are distance from the ground, worker's occupation and the source of the falling. The delivered severity prediction model provides the foundation of more accurate real time risk evaluation for workers at height.

Keywords: Falling from height; machine learning; serverity of injury; risk prediction

1. Introduction

Safety is always a key issue in the construction industry. Due to the dynamic and hazardous nature in construction jobsites, enhancing safety awareness and investigating the nature factors for accidents cannot draw more attentions. According to the statistics from OSHA, there are 13,344 fatality accidents during 2005 to 2015, and 35.1% accidents related to the construction industry [1]. Correspondingly, the HSE statistics data also shows that construction accidents has taken a large proportion [2]. This indicates that construction safety is still a challenging task. Considering the causes of worker deaths on construction sites, falls, struck by object, electrocutions and caught-in/between are the "Fatal Four" causes, which is 36.5%, 10.1%, 8.6% and 2.5% respectively [3].

As the most fatal cause in construction accidents, fall from height in construction worksite has drawn lots of researchers' attention. Lipscomb [4] assessed the rate of falls from height over the 20-year period from 1989 to 2008 by using the Poisson regression. They found that younger workers had higher injury rates; older workers lost more working days for fall accidents, and the rates of patterns of paid lost days associated with falls decreased over time. Especially for fatal falls from roofs in the U.S. construction industry, Dong [5] investigated the trends and patterns based on records from 1992 to 2009. The study shows that roof fatalities accounted for one-third of fatal falls in construction during 1992-2009, and 67% of deaths from roof falls occurred in small construction establishments. In the study, the workers younger than 20 years and older than 44 years have the relative higher risk of roof falling fatalities. Kaskutas [6] investigated 13 kinds of commercially available fall protection devices, and designed a survey to measure workers' perception for these protection devices. According to Kaskutas' study, it shows that many workers believe these technologies can help them prevent falls but decrease their productivity. Chi [7] studied the causes for the falls and the accident events, and analyzed the influences of fall sites, company sizes, causes of fall and individual factors on construction fall accidents. It shows that there are primarily three causes for fall accidents: lack of complying scaffolds or protections, improper use of PPE, and bodily actions.

* Corresponding author. Tel.: +852 3442 2971; fax: +852 3442 0427
*E-mail address:*xiaowluo@cityu.edu.hk

Considering the fall accidents, these researches focused more on the cause analysis, and they have investigated the influence of working environmental factors relationship and individual information on such construction safety incidents through statistical methods. According to their research, it has been proved that the objective factors of the construction companies, such as: the economic strength, project scale, etc. or the construction workers' individual factors, such as: gender, age, work experience, etc. definitely can affect the occurrence of the construction safety incidents. However, their research primarily relies on the qualitative analysis, the influences and the relations of the features and factors are not accurately quantized. In order to investigate the direct causes of the fall accidents on construction sites for proactive protection and early warning, more deeply data mining and correlation analysis are necessary.

In this paper, three kinds of data mining methods are employed (Decision-tree learning algorithm, Artificial Neural Network, and Clustering algorithm) to analyze the OSHA data. 15 OSHA recorded parameters are considered for investigating their influence on the injury severity level. Through the importance analysis for the parameters and comparison of 3 kinds of data mining algorithms, the features for proactive fall injury protection on construction sites has been analyzed. The rest of this paper is organized as follows: in Section 2, the data scheme, parameters value scale and three data mining methods are described; in section 3, the preliminary results are illustrated; and finally conclusions and future works are described in section 4.

2. Data Source and Methodology

2.1. *OSHA Data Statistics*

This research is based on the OSHA inspection data, and focuses on the fall accidences in the construction industry which the SIC Major Group codes include 15, 16, and 17. All the data since Jan. 01 1984 with specified 15 parameters have been obtained, which the EventType is fall, and any record which has none or blank value of parameters are filtered out, and finally there are 1166 intact records for this research.

Each record of an injury data is composed of 15 parameters, in which all the values of EventType parameter are fall. ID is the short sign of the corresponding parameter. Class is the type value of the parameter and the Code is the code of the corresponding type value. Number of Cases record the number of cases in totally 1161 records and the Frequency denotes the occurrence ratio. There are totally 13920 records which obtained from OSHA database, however, lots of the values of record parameters are blank or not reported. All the records which have blank or not reported parameter values are filtered, and then all the small probability recorded event, which the probability smaller than 5% is filtered. Finally, there are 1161 intact fall injury records.

2.2. *Clustering Algorithm*

Clustering is a direct method to classify the data set. In this research, classifying the recorded fall of the injury data is a primary way to analyze the fall injury accidents data. Clustering is a rather diverse topic, and the employed algorithms greatly rely on the application scenario [8]. In this study, most of the parameters of the data set are the type of categorical data; however, for most of the clustering algorithms, they primarily process the numerical data. Therefore in this study, the k-Modes is employed. k-Modes is a k-Means based algorithm which can be used the similarity measure instead of the distance measure, therefore especially be employed in categorical data clustering.

In k-Modes clustering algorithm, set $X = \{X_1, X_2, X_3,...X_n\}$ as a set of n categorical objects described by categorical attributes $A_1, A_2, ...A_m$. Then the mode of X is a vector Q that minimizes the following formula:

$$\mathrm{D}(X, Q) = \sum_{i=1}^{n} d(X_i, Q) \qquad (1)$$

where d is a similarity function which in this study it is the Hamming Distance function, and the mode X is also the cluster center. The process flow of the k-Modes algorithm is described as follow:

Step 1. Divide the data set into training data and test data according to the training sampling ratio.
Step 2. Set the number of target clusters n, according to the Degree values.
Step 3. Divide the training data set into n clusters, according the Degree values.
Step 4. Select the mode vectors for clusters, according to equation 1.
Step 5. Process test data, calculate the Hamming Distance between the test data and mode vector of each cluster. Then the test data belongs to the cluster, from which has the minimum Hamming Distance.
Step 6. Add the new classified data into the cluster, and re-calculate the mode vector of this cluster.
Step 7. Return to Step 5 until all the test data has been processed.

2.3. *Decision Tree Learning Algorithm*

Decision tree is a non-parametric classification and prediction model organized in the form of a rooted tree with at least 2 levels, and at least 2 branches at one or more levels that have two types of nodes called decision nodes and class nodes [9]. Predicting the value of a target variable through learning decision rules inferred from the data features is the goal of the Decision Trees algorithms. The major advantage of decision tree is that it performs well even when its assumptions are somewhat against the true model. And in addition, it requires little data preparation and can handle both numerical and categorical datas. In this research the CART (Classification and Regression Trees) based Decision Tree algorithm is employed. CART does not compute rule sets and can construct binary trees using the feature and threshold that yield the largest information gain at each node.

Set the training vectors $x_i \in R^n, i = 1,\ldots,I$, and the target label vector $y \in R^l$. A decision tree recursively partitions the training vectors space and divides the training vectors with the same labels into one group. Let Q presents the data set of node m, then the impurity at m can be calculated as $H()$, and the total impurity of the decision tree is $G()$.

$$p_{mk} = 1/N_m \sum_{x_i \in R_m} I(y_i = k)$$

$$H(X_m) = \sum_k p_{mk}(1 - p_{mk})$$

$$G(Q,\theta) = \frac{n_{left}}{N_m} H(Q_{left}(\theta)) + \frac{n_{right}}{N_m} H(Q_{right}(\theta))$$

The p_{mk} is the proportion of class k observations in node m. In this research, the Scikit-learn Decision Trees algorithm package is employed, and the decision tree prediction is implemented in Python.

2.4. *Artificial Neural Network Algorithm*

Artificial Neural Network (ANN) is a classical machine learning algorithm. In order to implement proactive protection for fall injury in the construction industry, to investigate the factors or rules related to the fall accidents is definitely essential. Different from Decision Tree learning algorithm, ANN algorithm is totally based on the training data instead of the information entropy, therefore the learning ratio of ANN algorithms can reflect the consistency of the training data. That means if ANN algorithms can learn the data set well, then there must be some direct and constant rules for the data set. In another word, the behavior which is reflected by the data set has strong predictability.

In this research, the RBF network is adopted to learn the training data set. RBF network mainly is used in classification, time series prediction, function approximation and system control. For RBF network, the input layer is a vector of real numbers $x \in R^n$, and the output of the network is a scalar function of the input vector $\varphi : R^n \to R$, therefore a RBF network can be descripted as follow:

$$\varphi(x) = \sum_{i=1}^{N} \partial_i \rho(\|x - c_i\|)$$

Where N presents the number of neurons of the hidden layer, c_i denotes the center vector for neuron i , and ∂_i is the weight of the output layer neuron i .

For this research, the intent is to predict the injury level according to the recorded accidents factors. Therefore, the output of the RBF network is the injury level which can be presented by the Degree parameters, and the input of the RBF network is the vector which consists of the recorded accident feature parameters. Since the output of the RBF network is the R domain, the value of the output may be floating numbers. However, in this research, the output is the category of the degree, and it is a kind of categorical data. In order to map the output value of RBF network into the given categorical value, k-Means algorithm is integrated to process the output value of RBF network.

3. Preliminary Results and Discussion

3.1. *Features Importance Analysis*

OSHA offers detailed information for construction fall accidents, including the basic description, occupational information, time and causes. All the recorded parameters can be quantized for this research. There are 15 parameters, in which the Degree presents the scale of the injury level. The recorded event type (EventType) is fall. Therefore, the Degree value is employed as the label of the data records in this research.

For investigating the influence of the parameters on the injury degree prediction. First, the importance of the parameters has been analyzed. Based on the Gini importance theory [10] and the mRMR algorithm [11], the importance of the parameters is calculated and the results are illustrated in Figure 1. Gini importance and mRMR are two classic features selection algorithms. These two algorithms can calculate the importance of features for classification, in which the larger importance value, the more influence for the classification.

As the illustrated results, the Occupation, FallDist and PartBody are the most 3 important parameters both in Gini importance analysis and the mRMR algorithm. For EventType parameter, it shows that the importance is 0 in the all situations. It is because that in this research the fall accidents are focused on, all the EventType values are fall. If just considering the injury degree, it shows that Occupation, FallDist, SourceInjury, and PartBody, the 4 parameters contribute most to the classification.

3.2. *Data Mining algorithms based Injury Severity Prediction*

Based on the OSHA data, one fall accident is an abstract record described by 15 features. According to the analysis in section 3.1, PartBody, FallDist, Occupation, and SourceInjury are the top 4 important features for classifying the data set into different injury degrees. The accuracy of the three types of classification algorithms has illustrated in Figure 2.

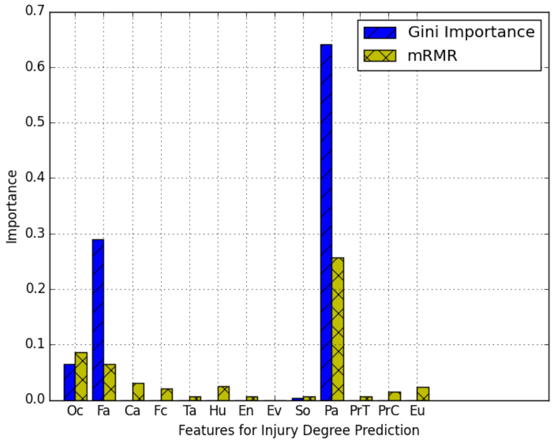

Figure 1. Features Importance Analysis for Injury Degree Prediction

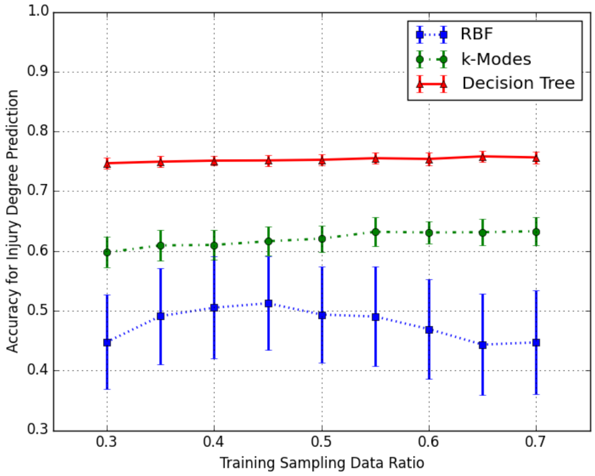

Figure 2. Training Data Ratio VS. Prediction Accuracy for Injury Degree

For the k-Modes algorithm, the value of the Degree feature is the label of the records and the k-Modes output. The value of the most relative features composed of the feature vectors and the input of k-Modes algorithm. Through learning the relation rules, the k-Modes algorithm predicts the injury degree according to the feature vectors. The k-Modes algorithm can generally achieve 0.62 accuracy.

RBF network also employed the most relative 4 features as the input vector, and the output is the degree of injury. Although generally RBF network is a good kind of ANN algorithm for classification, in this research, it can only achieve up to 0.49 accuracy for injury degree prediction.

Decision Tree algorithm based on the Gini importance theory to process the data vector. Based on the analysis in section 3.1, the top four important features were selected. Therefore the feature vectors can achieve the maximal Gini importance. Its accuracy for injury degree classification is around 0.75.

4. Summary and Conclusions

In order to investigate fall injury proactive protection for construction industry workers, this research studied the OSHA data. Two parameter analysis methods: Gini importance analysis and mRMR algorithm were employed to analyze the importance of the data features, and 3 kinds of data mining algorithms DT, k-Modes, and RBF network were applied to classify and predict the degree of the injury. By comparing the accuracy of the 3 kinds of algorithms, the DT algorithm definitely has the best performance for the injury prediction. According to the importance analysis, the injured part of the body (PartBody), fall down distance (FallDist), worker’s occupation (Occupations) and the source of the injury (SourceInjury) are the top 4 important parameters for injury degree prediction.

In conclusion, to a certain degree, the fall accidents in construction can be predicted, and the data mining algorithm, especially Decision Tree algorithm can do further contributions for fall injury proactive protection. In addition, according to the importance analysis and prediction comparison, just depending on the objective features of the construction company or construction workers to predict the construction injuries is definitely a reluctant method. In order to offer proactive protection, monitoring the position and posture of a worker, which can be relative to the FallDist and PartBOdy features, in real time is significant. This could be our future work.

Acknowledgements

This work was supported by National Science Foundation of China Grant # 51408519. The conclusions herein are those of the authors and do not necessarily reflect the views of the sponsoring agencies.

References

[1] OSHA, “Fatality and Catastrophe Investigation Summaries,” 2015. [Online]. Available: https://www.osha.gov/pls/imis/accidentsearch.html. [Accessed: 01-Mar-2016].

[2] D. Leigh, “Statistics on fatal injuries in the workplace in Great Britain 2015,” London, 2015.

[3] OSHA, “Commonly Used Statistics,” 2016. [Online]. Available: https://www.osha.gov/oshstats/commonstats.html. [Accessed: 01-Mar-2016].

[4] H. J. Lipscomb, A. L. Schoenfisch, W. Cameron, K. L. Kucera, D. Adams, and B. A. Silverstein, “How well are we controlling falls from height in construction? Experiences of union carpenters in Washington State, 1989-2008,” *Am. J. Ind. Med.*, vol. 57, no. 1, pp. 69–77, Jan. 2014.

[5] X. S. Dong, S. D. Choi, J. G. Borchardt, X. Wang, and J. A. Largay, “Fatal falls from roofs among U.S. construction workers,” *J. Safety Res.*, vol. 44, pp. 17–24, Feb. 2013.

[6] V. Kaskutas, B. A. Evanof, and H. Miller, “Fall prevention on residential construction sites,” *Prof. Safet*, pp. 36–40.

[7] C.-F. Chi, T.-C. Chang, and H.-I. Ting, “Accident patterns and prevention measures for fatal occupational falls in the construction industry,” *Appl. Ergon.*, vol. 36, no. 4, pp. 391–400, Jul. 2005.

[8] C. C. Aggarwal and C. K. Reddy, *Data Clustering: Algorithms and Applications*. Chapman and Hall/CRC, 2013.

[9] C.-W. Liao and Y.-H. Perng, “Data mining for occupational injuries in the Taiwan construction industry,” *Saf. Sci.*, vol. 46, no. 7, pp. 1091–1102, Aug. 2008.

[10] L. Breiman, J. Friedman, C. J. Stone, and R. A. Olshen, *Classification and Regression Trees*. Chapman and Hall/CRC, 1984.

[11] Hanchuan Peng, Fuhui Long, and C. Ding, “Feature selection based on mutual information criteria of max-dependency, max-relevance, and min-redundancy,” *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 27, no. 8, pp. 1226–1238, Aug. 2005.