Journal of Safety Research 58 (2016) 79–87

Contents lists available at ScienceDirect

Journal of Safety Research

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

Off-road truck-related accidents in U.S. mines

Saeid R. Dindarloo, [a,*] Jonisha P. Pollard, [b] Elnaz Siami-Irdemoosa [c]

[a] Department of Mining and Nuclear Engineering, Missouri University of Science and Technology, MO, USA
[b] Workplace Health Branch, Pittsburgh Mining Research Division, NIOSH, PA, USA
[c] Department of Geoscience and Geological and Petroleum Engineering, Missouri University of Science and Technology, MO, USA

ARTICLE INFO

Article history:
Received 17 January 2016
Received in revised form 2 May 2016
Accepted 12 July 2016
Available online 25 July 2016

Keywords:
Off-road mining trucks
Fatalities and injuries
K-means clustering
Genetic programming
Classification

ABSTRACT

Introduction: Off-road trucks are one of the major sources of equipment-related accidents in the U.S. mining industries. A systematic analysis of all off-road truck-related accidents, injuries, and illnesses, which are reported and published by the Mine Safety and Health Administration (MSHA), is expected to provide practical insights for identifying the accident patterns and trends in the available raw database. Therefore, appropriate safety management measures can be administered and implemented based on these accident patterns/trends. *Methods:* A hybrid clustering-classification methodology using K-means clustering and gene expression programming (GEP) is proposed for the analysis of severe and non-severe off-road truck-related injuries at U.S. mines. Using the GEP sub-model, a small subset of the 36 recorded attributes was found to be correlated to the severity level. *Results:* Given the set of specified attributes, the clustering sub-model was able to cluster the accident records into 5 distinct groups. For instance, the first cluster contained accidents related to minerals processing mills and coal preparation plants (91%). More than two-thirds of the victims in this cluster had less than 5 years of job experience. This cluster was associated with the highest percentage of severe injuries (22 severe accidents, 3.4%). Almost 50% of all accidents in this cluster occurred at stone operations. Similarly, the other four clusters were characterized to highlight important patterns that can be used to determine areas of focus for safety initiatives. *Conclusions:* The identified clusters of accidents may play a vital role in the prevention of severe injuries in mining. Further research into the cluster attributes and identified patterns will be necessary to determine how these factors can be mitigated to reduce the risk of severe injuries. *Practical application:* Analyzing injury data using data mining techniques provides some insight into attributes that are associated with high accuracies for predicting injury severity.

1. Introduction

Analysis of workplace injuries has been heavily utilized as a means to determine high-risk tasks, prioritize workplace redesign, and determine areas of concern for worker safety in many industries including healthcare, construction, retail and services, and mining (Cato, Olson, & Studer, 1989; Drury, Porter, & Dempsey, 2012; Mardis & Pratt, 2003; Moore, Porter, & Dempsey, 2009; Pollard, Heberger, & Dempsey, 2014; Schoenfisch, Lipscomb, Shishlov, & Myers, 2010; Turin, Wiehagen, Jaspal, & Mayton, 2001; Wiehagen, Mayton, Jaspal, & Turin, 2001). While many industries would require injury records from individual companies or insurance providers to perform an analysis, mining is uniquely suited for a more comprehensive injury analysis. An important feature of U.S. mining is the accessibility of injury records. The Mine Safety and Health Administration requires all mine operators and contractors to file a Mine Accident, Injury and Illness Report (MSHA Form 7000-1) for all reportable accidents, injuries, and illnesses incurred at U.S. mining facilities. Reportable illnesses include any illness or disease that may have resulted from work. The database of these reports is available in the public domain and is provided by the National Institute for Occupational Safety and Health (http://www.cdc.gov/niosh/mining/data/default.html). Each entry of the database contains 36 unique attributes including: mine id, mining method, accident date, degree of injury, accident classification, mining equipment, employee's experience and activity, and a narrative briefly explaining the accident. Previous mining research has examined the injury and fatality causes associated with maintenance and repair, haulage vehicles, ingress and egress from mobile equipment, operating underground and surface mining mobile equipment, and other mining tasks (Drury et al., 2012; Moore et al., 2009; Pollard et al., 2014; Reardon, Heberger, & Dempsey, 2014; Turin et al., 2001; Wiehagen et al., 2001). Traditional injury data analysis uses counts and cross-tabulations as a means to determine trends in injuries. While this typically yields useful information, more sophisticated data mining techniques may allow for more improved classification of injuries through identification of injury patterns.

Clustering and classification are the two widely used methods of data mining for the purpose of pattern recognition. Clustering is

* Corresponding author at: 226 McNutt Hall, 1400 N. Bishop Ave., Rolla, MO 65409, USA.
E-mail address: srd5zb@mst.edy (S.R. Dindarloo).

http://dx.doi.org/10.1016/j.jsr.2016.07.002

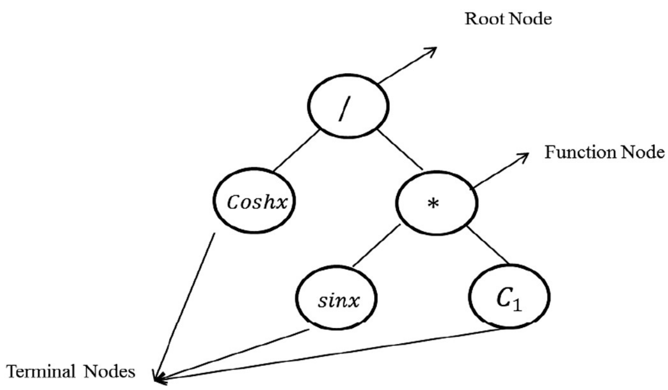

Fig. 1. GP tree representation of $Coshx/(C_1 sinx.)$.

among the unsupervised methods of pattern recognition while the classification is a supervised learning method. By an unsupervised method, one means that the data analyzer does not have any prior hypothesis or pre-specified models for the data, but wants to understand the general characteristics or the structure of the high-dimensional data. A supervised method means that the investigator wants to confirm the validity of a hypothesis/model or a set of assumptions, given the available data (Jain, 2010). Clustering and classification are also called un-labeled and labeled, respectively. In pattern recognition, data analysis is concerned with predictive modeling: given some training data, the prediction task is to find the behavior of the unseen test data. This task is also referred to as learning. Often, a clear distinction is made between learning problems that are (i) supervised (classification) or (ii) unsupervised (clustering), the first involving only labeled data (training patterns with known category labels), while the latter involves only unlabeled data (Duda, Hart, & Stork, 2001; Jain, 2010). Clustering and classifications are performed using differing algorithms but may be used together to improve prediction accuracy.

The aim of this research was to gain a better understanding of the factors associating with severe injuries (fatalities and permanent disabilities) in U.S. mining by employing data mining techniques. Clustering and classification were employed for a comprehensive analysis of off-road truck-related accidents and injuries reported to MSHA during a 13-year period (2000–2012). Gene expression programming was used for classification, allowing all injury attributes to be considered and tested to determine which were associated with the highest prediction accuracies. The most explanatory attributes were selected among the available 36 unique attributes in the MSHA database. Then, K-means clustering was used as a means to identify similarity/dissimilarity between the accident records using the selected attributes for the purposes of pattern recognition in the raw data. It should be noted that the goal of this study was not to establish cause–effect relationships between accident attributes and outcomes, but to: (a) use data mining to systematically identify important attributes from MSHA incident reports that are highly associated with the outcomes of accidents (classification), and (b) recognize patterns in the accidents (clustering) given a set of work-related attributes.

2. Materials and methods

2.1. MSHA injury data

A dataset comprised of 13 years (2000–2012) of Mine Accident, Injury and Illness Reports was selected beginning with 1/1/2000 (MSHA, 2014). From this dataset, records of severe injuries (fatalities and permanent disabilities) and non-severe injuries associated with off-road trucks were selected. The NIOSH code “minemach-44, all accidents related to off-road mining trucks” was identified to select the records of interest in this study. A total of 5,831 records of injuries (both severe and non-severe) were filtered for further analysis. This dataset included 125 severe records that affected 140 employees. These severe injuries consisted of 88 fatalities and 52 permanent disabilities. These records were analyzed using Minitab (Minitab Inc., State College, Pennsylvania), MATLAB (MathWorks, Inc., Natick, Massachusetts), Rapidminer (RapidMiner, Inc., Cambridge, Massachusetts) and GenExprotools (Gepsoft Limited, Bristol, UK) to determine factors associated with the highest counts of severe injuries.

2.2. GEP-clustering modeling

The objective of clustering is to discriminate between dissimilar data by dataset partitioning (clustering). As an unsupervised data mining technique, the aim of clustering is to split a heterogeneous dataset into several more homogenous groups. The optimization task is to maximize the similarity between the in-cluster members and dissimilarity between the out-cluster members. K-means clustering is used to partition a large, highly variable dataset such that like data are grouped together. As an example, one is given a set of n data points in d-dimensional space (R^d) and an integer k. The goal is to determine a set of k points in R^d, called centers, so as to minimize the mean squared distance from each data point to its nearest center (Kanungo et al., 2002). Let $X = \{x_i\}, i = 1, \ldots, n$,be the d-dimensional observations which are clustered into a set of k clusters, $C = \{c_k, K = 1, \ldots, K\}$. The K-means clustering finds a partition such that the squared error between the empirical mean of a cluster and the points in the cluster is minimized. Let μ_k be the mean of the cluster c_k. The squared error between μ_k and the points in cluster c_k is defined as shown in Eq. (1). The goal of K-means clustering is to minimize the sum of the squared error over all k clusters as shown in Eq. (2). (For a detailed review of the theory and background of K-means clustering see: Halkidi et al., 2001, Fraley and Raftery, 2002, Jain, 2010.)

$$J(c_k) = \sum_{x_i \in c_k} \|x_i - \mu_k\|^2 \tag{1}$$

$$J(c) = \sum_{k=1}^{K} \sum_{x_i \in c_k} \|x_i - \mu_k\|^2 \tag{2}$$

Genetic programming (GP) creates a functional relationship between inputs (attributes) and outputs to predict the occurrence of the output based on the properties of the attributes. Genetic programming can be represented as a hierarchically structured tree comprising functions and terminals. Fig. 1 illustrates a simple representation of a GP tree for the function $Coshx/(C_1 sinx)$. The tree reads from left to right and from bottom to top. Mimicking the Darwinian principle of survival, the fittest solutions (smallest error) are chosen to generate a population of new offspring programs for the next generation (Koza, 1992). In the

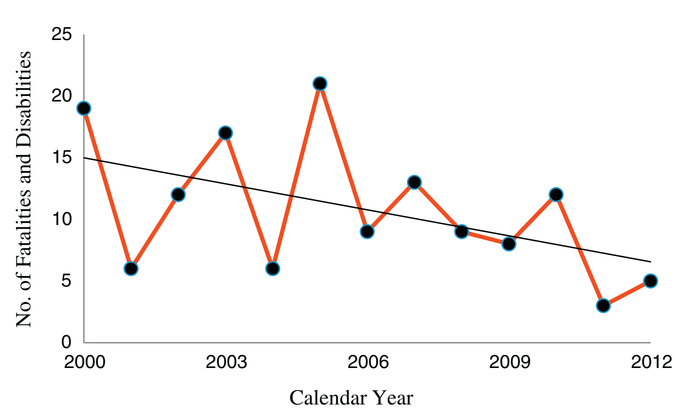

Fig. 2. Time series of off-road truck-related severe injuries at US mines (2000–2012).