Abstract

Name of the student: **Sobhan Sarkar** Roll No: **14IM91R03**
Degree for which submitted: **Doctor of Philosophy (Ph.D.)**
Department: **Department of Industrial & Systems Engineering**
Thesis title: **PREDICTIVE MODELING OF OCCUPATIONAL INCIDENTS USING MACHINE LEARNING TECHNIQUES**
Thesis supervisor(s): **J Maiti**
Month and year of thesis submission: **November, 2019**

Although the usefulness of machine learning (ML) techniques in predicting future outcomes has been established in different domains of applications (e.g., healthcare), its exploration in the prediction of incidents in occupational safety domain is new. This necessitates the investigation of ML techniques in this research domain. This research aims at the development of a comprehensive methodology for predictive modeling of occupational incidents. The thesis addresses five important objectives as (i) to develop a methodology for prediction of occupational incident outcomes using both structured and unstructured data, (ii) to develop a methodology for rule-based pattern extraction of the causal factors of incidents, (iii) to develop a methodology for prediction of occupational injury risk, (iv) to develop a methodology for prediction of injury severity using both proactive and reactive data, and (v) to develop a decision support system (DSS) for prediction of occupational incident.

The first objective is addressed using both categorical and text data. Latent Dirichlet Allocation (LDA)-based topic modeling is used for handling text data. For prediction, genetic algorithm (GA), and particle swarm optimization (PSO)-based support vector machine (SVM) and artificial neural network (ANN) are used. The second objective is addressed using GA and PSO-based decision tree algorithms, namely C5.0, classification and regression tree analysis (CART), and random forest (RF). The third objective is addressed using GA, PSO, artificial bee colony (ABC) and Firefly algorithm (FA)-based artificial neuro-fuzzy inference system (ANFIS) for the prediction of injury risk. To reduce the redundancies

and inconsistencies in data, rough set theory (RST) is used in data pre-processing step. Finally, RST is employed to extract a set of crisp rules for injury risk. The fourth objective is addressed using reactive and proactive data simultaneously. Class imbalance issue is handled using oversampling techniques, namely Synthetic Minority Over-sampling Technique (SMOTE), Borderline SMOTE (BLSMOTE), Majority Weighted Minority Oversampling Technique (MWMOTE), and k-means SMOTE (KMSMOTE), separately. For prediction, a set of five classification algorithms, namely SVM, K-nearest neighbour (KNN), Naive Bayes (NB), CART, and RF are used, separately. Tolerance rough set approach (TRSA) is used to extract a set of crisp rules for injury severity. The fifth and final objective is addressed by developing a decision support system, named as 'AcciDSS' for the prediction of occupational incident. To validate the developed AcciDSS, a case study of prediction of STF occurrences and extraction of safety decision rules for these occurrences is presented.

The thesis has made contributions from theoretical development and practical implication points of view. From the theoretical development point of view, the thesis contributes as follows: (i) a methodology is developed to effectively consider both unstructured texts and structured categorical data for prediction; (ii) a rough set theory (RST)-based prediction model developed for handling redundancies and inconsistencies in data; (iii) a methodology is developed for handling class imbalance and missing values in data for prediction.; (iv) a methodology is developed for parameter optimization of the classifiers using optimization algorithms to obtain improved prediction performance; (v) a rule-based prediction model is developed for analysing causal factors behind any incident; (vi) a methodology is developed to extract a set of crisp safety decision rules using RST and tolerance rough set approach. On the other hand, from the practical implication point of view, the thesis contributes as follows: (i) a methodology is developed for prediction of injury severity using both reactive and proactive data; (ii) a DSS is developed for the purpose of practical implementation through predicting and analysing occupational incidents. Using the incident data collected from an integrated steel plant located in India, the methodologies presented in this thesis are validated.

From the analyses presented in the thesis, some key findings are obtained. They are: (i) the use of both text and categorical data helps to achieve better prediction accuracy than using only categorical data in prediction of occupational incident outcomes; (ii) parameter optimization of classification algorithms is found to be very useful; (iii) PSO-based SVM outperforms other algorithms with 90.67% accuracy in predicting incident outcomes; (iv) PSO-based RF is found to be the best classifier with 78.819% accuracy in predicting STF and generating a set of 20 interpretable safety decision rules explaining the factors behind

the occurrences of STFs; (v) PSO-ANFIS algorithm outperforms other algorithms in the prediction of injury risk with mean absolute error (MAE) 14.2% for training and 17.5% on testing dataset; (vi) RST-based approach for handling data redundancies and inconsistencies during data pre-processing helps to achieve increased prediction; (vi) prediction of injury severity using both reactive and proactive data is found to be more effective than using only reactive data; (vii) KMSMOTE algorithm is observed to be very effective oversampling technique for handling class imbalance issue in data; and (viii) the developed DSS can be used in data pre-processing, descriptive & predictive analysis, and rule-based pattern extraction of incidents in incident analysis and prevention.

Keywords: Occupational incident, Predictive modeling, Machine learning, Incident outcome prediction, Rule-based decision making, Injury risk prediction, Injury severity prediction, Reactive and proactive data, Class imbalance, Decision support system