Application of Data-Mining Techniques for Personal Injury Evaluation in Tanker Shipping Industry

Qingji Zhou[1], Vinh V. Thai[1,2]

***Abstract*—**Both adaptive neuro-fuzzy inference system (ANFIS) and other data-mining methods were applied to study the personal injury on oil tankers using database of personal injury accidents recorded by tanker shipping companies. Chi-square test method was applied to determine the most important variables causing the accidents. The selected variables proved to be correlated to the observed frequency of four injury categories, namely FAC (First Aid Case), LTI (Lost Time Injury), MTC (Medical Treatment Case) and RWC (Restricted Workday Case). A Gravity Factor (GF) was calculated based on the percentage of injury categories resulting to the accidents for each variable determined by Chi-square test. The calculated GF values and the severity of the accidents were used as input values in ANFIS model. For ANFIS, trapezoidal and gauss membership functions were used, both fuzzy logical theory and artificial neural networks were applied for the training of the data by MATLAB software. The machine learning algorithms in WEKA software were also used. Both of the predictive results were compared with the original recorded data for verifying these methods. The methods and the results also can be used in decision-making and suggesting optimal risk control measures, and will be of significant managerial contributions to the safety of tanker shipping industry.

***Keywords*—** ANFIS, machine learning algorithm, Chi-square test, WEKA.

I. Introduction

IN recent years, the growth rate of world oil and gas demand is raised rapidly. According to BP statistical review of world energy 2015, the oil and natural gas consumption is 3870.8 million tons and 2711.3 billion cubic meters in 2004, up to 4211.1 million tons and 3460.6 billion cubic meters respectively in 2014[1]. According to the world energy outlook forecast by International Energy Agency (IEA), the world's oil and gas demand is expected to reach 5.769 billion tons and 4.203 billion tons of oil equivalents respectively in 2030. Many, at that time and still, believe oil tanker is a specialized ship designed to transport oil from offshore oil fields to onshore refineries. Also tankers are often used as an alternative to pipelines in harsh climates, remote locations or deep-water area. From 1970 to 2013, international tanker cargo trade increased from 1440 million tons to 2844 million tons.

It is a well-known fact that shipping and ship management companies face lots of potential hazards in relation to occupational health and safety for their crew, vessel safety and security, and environmental (HSSE) risks in their daily operations. People working on the tankers not only face the response to the motion of the ships, the fatigue of the crew on the ships, the toxic risk especially during tank cleaning, loading and unloading of gasoline, they will also face all kinds of potential personal injuries during different kinds of operations. They will get hurt in any part of the body even fatally. Shipping accidents prevention is quite a, common topic of research which attracts the attention of various scholars. Forecasting the occupational injury on tankers will be of significant contribution to the safety of tanker shipping industry. Data-mining techniques, which have been used in many industries, obviously are proper to evaluate the severity of personal injury on board.

II. Literature review

Data-mining techniques, including decision rules, classification trees, Bayesian networks, support vector machines and logistic regression, have been applied in occupational injury forecasting for several years[2][3].

Bevilacqua et al (2008)[4] applied classification tree methods to data regarding accidents in a medium-sized refinery, so as to identify the important relationships between the variables, which can be considered as decision-making rules when adopting any measures for improvement. A methodology for the analysis of the causes and types of workplace accidents had been proposed by MATÍAS et al (2006)[5], the approach was based on machine learning techniques, Bayesian networks trained using different algorithms (with and without a priori information), classification trees, support vector machines and extreme learning machines. Nenonen (2013)[6] applied methods of data mining (decision tree and association rules) to the Finnish national occupational accidents and diseases statistics database to analyse factors related to slipping, stumbling, and

[1] Maritime Institute@NTU, Nanyang Technological University, Nanyang Avenue, Singapore

[2] School of Civil & Environmental Engineering, Nanyang Technological University, Singapore.

falling accidents at work, proving data mining methods were seen as a useful supplementary method in analysing occupational accident data. Cheng et al (2012)[7] explored the causes and distribution of occupational accidents in the Taiwan construction industry by analysing such a database using the data mining method known as classification and regression tree (CART), the results of the study provided a framework for improving the safety practices and training programs that were essential to protecting construction workers from occasional or unexpected accident. BEVILACQUA et al (2010)[8] applied data mining techniques to data regarding accidents in a medium-sized refinery, the results indicated important relationships between the variables, providing useful decision-making rules which can be followed when adopting measures for improvement. Rivas et al (2011)[9] introduced data-mining techniques to model accident and incident data compiled from the mining and construction sectors, the results were compared with those for a classical statistical techniques (logistic regression), revealing the superiority of decision rules, classification trees and Bayesian networks in predicting and identifying the factors underlying accidents. An adaptive neuro-fuzzy inference system (ANFIS) had been applied to study the effect of working conditions on occupational injury using data of professional accidents assembled by ship repair yards by Fragiadakis et al (2014)[10].

The focus of this research is to create an effective method for the personal injury evaluation in tanker shipping industry. Data-mining methods have been applied to study the effect of different variables on the personal injury. Software computing techniques are applied to estimate the personal injury on tankers. The methods and the results also can be used in decision-making and suggesting optimal risk control measures, and will be of significant managerial contribution to the safety of tanker shipping industry.

III. METHODOLOGY

A. Variables selection

There are 285 personal injury accidents recorded by a global tanker ship management company from 2008 to March 2015. According to the investigation reports, there are many variables affecting the personal injury on board, such as vessel type, vessel team, vessel age, seafarer information (age, nationality, rank, time on board, time in rank, time in company), hurt category, place, trading area, site of the tanker, operation and so on. First of all, correlation analysis method (Chi-square test) should be applied to determine the variables that have strong relationship with the accident. Compared to personal injury severity, the results of P-value of the variables are shown in TABLE I.

TABLE I: P-VALUE OF THE VARIABLES COMPARED TO PERSONAL INJURY SEVERITY

Variables	P-value	Variables	P-value	Variables	P-value
Nationality	0.0432	Time in rank	0.5424	Location	0.0043
Age	0.3328	Vessel type	0.0478	Site	0.6040
Time on board	0.2278	Vessel age	0.9620	Operation	0.0288
Time in company	0.1298	Vessel team	0.5464	Trading area	0.0115
Rank	0.3830				

Usually, the standard level of significance used to justify a claim of a statistically significant effect is 0.05. For better or worse, the term statistically significant has become synonymous with P less or equal to 0.05. Therefore, the injury frequency index is produced taking into consideration of nationality, vessel type, location, operation and trading area. The personal injury severity on tankers includes FAC (First Aid Case), LTI (Lost Time Injury), MTC (Medical Treatment Case) and RWC (Restricted Workday Case). ANFIS and other data-mining methods will be applied to evaluate personal injury on the tanker.

B. ANFIS structure and results

The adaptive network based fuzzy inference system (ANFIS) is a kind of artificial neural network that is based on Takagi–Sugeno fuzzy inference system [11]. Both artificial neural network and fuzzy logic are used in ANFIS architecture. An ANFIS can construct an input-output data pair for neural networks training by using hybrid learning method. The ANFIS structure is used for training Sugeno type FIS through learning and adaptation. Usually there are five layers in the structure, fuzzification layer, product layer, normalized layer, de-fuzzification layer and output layer. ANFIS requires a training data set of desired input/output pair (x1,x2,...,xm,y) depicting the target system to be modelled. ANFIS adaptively maps the inputs (x1,x2,...,xm) to the output y through MFs (membership functions), the rule base, and the related parameters emulating the given training dataset.

According to Fragiadakis et al (2014), the frequency can be calculated as the number of times the same level of injury has occurred under the same accident factors. The above-mentioned parameters are codified according to the accident records. This codification results in a specific number of categories for each variable. All the above mentioned categories are presented in TABLE II. A Gravity Factor (GF) was produced from the injury frequency values for each variable and used as input value in order to evaluate the severity. A normalization formula was used to estimate the Gravity Factor (GF) of each category concerning the factors in TABLE III.

TABLE II: GF VALUES OF THE CATEGORIES IN EACH VARIABLE

Variables	Categories	GF	Variables	Categories	GF
Nationality	Malaysian	0.3188	Trading area	Worldwide	0.3672
	Indian	0.2897		Asia	0.2922
	Filipino	0.3333		US Gulf	0.3800
	Other countries	0.3303		US/Canada	0.6667
Vessel type	AFRA	0.2793		Europe	0.4667
	VLCC	0.3788	Operation	Machinery	0.4000
	Others	0.4065		Maintenance	0.3867
Location of tanker	At Sea	0.3153		Routine work	0.2121
	Anchorage	0.2544		Handling a load	0.5333
	In Port	0.3750		Using Tools	0.1795
	Offshore	0.6667		Stair	0.1778
	Shipyard	0.5556		Cleaning	0.4222
	Shipyard Dry-dock	0.0833		Mooring	0.3016
				Others	0.2869

TABLE III: PERSONAL INJURY SEVERITY VALUE

i	Normalization factor	Range of GF	Personal injury severity
1	0	0.00-0.25	FAC
2	1	0.25-0.50	LTI
3	2	0.50-0.75	MTC
4	3	0.75-1.00	RWC

The GF can be calculated by:

$$GF = \frac{\sum x_i y_i}{(n-1)\sum x_i} \quad (1)$$

Where n is the categories of risk level, for personal injury accidents, there are 4 consequences, so n=4; where xi (i=1,…,n) is the percentage for each resulting occupational injury and yi is the respective normalization factor for risk value calculation according to TABLE III. Thus the final value of GF is scaled from 0 to 1. The resulting values for each categories of each variable are presented in TABLE II. These values of GF were finally used as input values in the ANFIS.

A total number of 158 data sets were obtained in the statistical procedure, by removing some data missing samples. Most of the data sets were selected for training; only 20 samples were used for testing after training was completed to verify the accuracy of the predicted results.

TABLE IV: PARTS OF INPUT DATA SET SAMPLES USED FOR TRAINING TO DEVELOP THE MODEL

Data no.	Nationality	Vessel type	Location of tanker	Trading area	Operation	Personal injury severity
1	0.3188	0.3788	0.375	0.2922	0.2869	0.500
2	0.2897	0.2793	0.3153	0.2922	0.3867	0.250
3	0.3333	0.2793	0.2544	0.38	0.2869	0.250
4	0.3333	0.2793	0.2544	0.38	0.2869	0.250
5	0.2897	0.2793	0.2544	0.38	0.5333	0.500
6	0.2897	0.2793	0.3153	0.3672	0.2121	0.250
7	0.3188	0.2793	0.2544	0.38	0.2869	0.250
8	0.2897	0.4065	0.3153	0.3672	0.3867	0.250
9	0.2897	0.2793	0.3153	0.3672	0.4222	1.000
10	0.2897	0.2793	0.2544	0.3672	0.2121	0.250
11	0.3333	0.2793	0.375	0.38	0.3016	0.250
12	0.2897	0.3788	0.3153	0.2922	0.2121	0.250
13	0.2897	0.2793	0.375	0.2922	0.2869	0.750
14	0.2897	0.3788	0.375	0.2922	0.4	0.500
15	0.2897	0.2793	0.2544	0.38	0.2869	0.250

For evaluating the risk in tanker shipping, an input/output data set was applied to construct a fuzzy inference system, two types of membership functions were used, the Trapezoidal type and the Gaussian type membership function. For each type of membership function three input membership functions were used for each variable category. This allows fuzzy systems to learn from the data they are modelling. ANFIS applies two techniques in updating parameters. For premise parameters that define membership functions, ANFIS employs gradient descend to fine-tune them. For consequent parameters that define the coefficients of each output equations, ANFIS uses the least-squares method to identify them. This approach is thus called hybrid learning method since it combines the gradient descend method and the least-squares method.

TABLE V: RESULTING PERSONAL INJURY SEVERITY MEASURED AND PREDICTED FROM ANFIS TRAINING

Data no.	Nationality	Vessel type	Location of tanker	Trading area	Operation	personal injury severity	Predicted severity Trapezoidal MF	Error	Predicted severity Gauss MF	Error
1	0.3188	0.2793	0.3750	0.3800	0.5333	0.7500	0.519	0.231	0.816	0.066
2	0.2897	0.3788	0.3750	0.3672	0.2869	0.2500	0.458	0.208	0.439	0.189
3	0.2897	0.2793	0.2544	0.3800	0.2869	0.7500	0.409	0.341	0.415	0.335
4	0.3333	0.2793	0.3750	0.3672	0.3016	0.2500	0.323	0.073	0.355	0.105
5	0.2897	0.2793	0.3153	0.2922	0.1795	0.2500	0.247	0.003	0.359	0.109

6	0.3188	0.3788	0.3153	0.3672	0.1778	0.2500	0.643	0.393	0.615	0.365
7	0.2897	0.3788	0.3750	0.3800	0.1778	0.2500	0.090	0.16	0.180	0.07
8	0.2897	0.2793	0.2544	0.3800	0.3016	0.5000	0.445	0.055	0.464	0.036
9	0.2897	0.2793	0.2544	0.3800	0.5333	0.2500	0.506	0.256	0.500	0.25
10	0.2897	0.3788	0.3153	0.3672	0.2869	0.7500	0.500	0.25	0.500	0.25
11	0.2897	0.2793	0.2544	0.3672	0.2121	0.2500	0.228	0.022	0.218	0.032
12	0.3188	0.2793	0.2544	0.2922	0.2869	0.5000	0.252	0.248	0.252	0.248
13	0.3188	0.2793	0.3750	0.2922	0.3016	0.7500	0.421	0.329	0.338	0.412
14	0.3333	0.3788	0.2544	0.3672	0.4000	0.5000	0.696	0.196	0.291	0.209
15	0.2897	0.2793	0.2544	0.3800	0.3867	0.7500	0.605	0.145	0.641	0.109
16	0.3188	0.2793	0.3153	0.2922	0.2869	0.5000	0.334	0.166	0.546	0.046
17	0.2897	0.2793	0.3153	0.3672	0.2869	0.2500	0.390	0.14	0.285	0.035
18	0.2897	0.4065	0.3153	0.3672	0.2869	1.0000	0.558	0.442	0.711	0.289
19	0.2897	0.3788	0.3153	0.3672	0.2869	0.7500	0.500	0.25	0.500	0.25
20	0.2897	0.2793	0.2544	0.3800	0.3867	0.7500	0.605	0.145	0.641	0.109
						Average error		0.176		0.153

In TABLE V, comparison of the predicted severity and original value is presented. Because the quantity of the recorded accidents is not large, the average error of the personal injury severity prediction is around 17.6% by Trapezoidal type membership function and 15.3% by Gaussian type membership function. The prediction accuracy of using Gaussian membership function is higher than that when the Trapezoidal membership function is used. Gaussian type membership functions, compared with other types, have the advantage of having a concise notation as well as other useful properties such as invariance in multiplication and the fact that the Fourier transform of a Gaussian function is another Gaussian.

C. Waikato Environment for Knowledge Analysis (WEKA)

WEKA (Waikato Environment for Knowledge Analysis) is a popular suite of machine learning software written in Java, developed at the University of Waikato[12], New Zealand. WEKA is free software available under the General Public License. It is a workbench that contains a collection of visualization tools and algorithms for data analysis and predictive modelling, together with graphical user interfaces for easy access to these functions.

The data-mining techniques used to model the personal injury severity were classification trees, Bayesian networks, decision rules, support vector machines and logistic regression[9].

Classification trees are statistical classification techniques that can be graphically represented as diagrams. There are different kinds of trees, but they are all generally trained by progressively dividing the data into groups. Each group is as similar as possible in terms of the response variable. Each group obtained in the previous stage is divided again, with a view to enhancing similarity, using a new condition based on an explanatory variable, and so on successively until some stop criterion is satisfied[13][14]. The algorithm of classification trees includes ID3 algorithm, J48 algorithm, LMT algorithm and so on.

Bayesian networks are directed acyclic graphs used for descriptive and predictive purposes. In this paper, K2 and hill-climber implemented in WEKA are used as the network training algorithms, with different constraints on the number of parents. Meanwhile, naive Bayes, with a structure of just two levels and a single parent is applied. The networks were trained by means of a greedy search of the space of possible structures, with the best network chosen on the basis of a specific goodness-of-fit criterion for the selected algorithm[15].

Logistic regression can be seen as a special case of generalized linear model and thus analogous to linear regression. The model of logistic regression, however, is based on quite different assumptions (about the relationship between dependent and independent variables) from those of linear regression, which represents the probability of occurrence of the class of interest (accident) versus the probability of the occurrence of another class (incident). Logistic regression is estimated using the maximum likelihood method[16].

Decision rule is a function which maps an observation to an appropriate action. Decision rules play an important role in the theory of statistics and economics, and are closely related to the concept of a strategy in game theory[17].

TABLE VI: VARIABLES AND PERCENTAGE OF THE CATEGORIES USED IN WEKA

Variables	Categories	Frequency	Percentage (%)
Nationality	Malaysian	40	25.32
	Indian	78	49.37
	Other countries	40	25.32
Vessel type	AFRA	109	68.99
	VLCC	33	20.89
	Others	16	10.13
Location of tanker	At Sea	66	41.77
	Anchorage	48	30.38
	In Port	38	24.05
	Offshore	2	1.27
	Shipyard	3	1.90
	Shipyard-Dry-dock	1	0.63
Trading area	Worldwide	47	29.75
	Asia	67	42.41
	US Gulf	36	22.78
	US/Canada	7	4.43
	Europe	1	0.63
Operation	Machinery	14	8.86
	Maintenance	25	15.82
	Routine work	15	9.49
	Handling a load	9	5.70
	Using Tools	12	7.59
	Stair	11	6.96
	Cleaning	11	6.96
	Mooring	10	6.33
	Others	51	32.28

Source: Database of the global tanker ship management company

The data are both used for training and testing. The results can be seen in TABLE VII. Compare to existing literature, the success rates for forecasting the personal injury severity were not high, mainly because the recorded samples are not large enough. As for success rates, the best results were obtained by the decision rule with part algorithm (74.21), the classification tree with the J48 tree algorithm (70.44%), the Logistic regression (64.15%). The results of Bayesian network methods were very similar, around 60%.

TABLE VII: SUCCESS RATES FOR FORECASTING THE PERSONAL INJURY SEVERITY BY WEKA

Model	Success (%)
BayesNet—K2—1 parent	60.38
BayesNet—K2—3 parents	59.75
BayesNet—HC—1 parent	61.01
BayesNet—HC—3 parents	57.86
Naive Bayes	58.49
Simple Naive Bayes	58.49
Logistic regression	64.15
Tree—J48	70.44
Tree—LMT	53.46
Rule—PART	74.21
Rule—OneR	54.09

IV. CONCLUSION

The method in this paper can be used for handling and analysing workplace accident data that identifies the most relevant variables and consequently improves prediction success rates and explanatory capacities. In this research, ANFIS, Bayesian network, logistic regression, classification and tree decision rule methods were used to evaluate the personal injury severity in tanker shipping industry using the data of professional accidents assembled by tanker shipping company. The data were statistically processed firstly by Chi-square test in order to get the most important variables, which were nationality, vessel type, location, operation and trading area. A Gravity Factor (GF) was calculated, by comparing the percentage of the categories in each variable to the severity of the accidents. These GF values and the resulting severity value based on the accident data were used as inputs for training in ANFIS method.

WEKA software was applied to evaluate personal injury severity on board. The results of WEKA analysis represent an important advance in terms of managing information on personal injury accidents on tankers. The machine learning algorithms used in this paper seem to be proper tools for the studies of personal injury accidents on board. The quality of the results can be applied to gain a deeper understanding of the cause of the accidents.

The accuracy of ANFIS and other data-mining methods were not so high in this paper, this mainly because the quantity of the recorded samples used for training and classification were not large enough, and the severity of personal injury has related with several variables and many categories. But still, the results can be applied in decision-making and suggesting optimal risk control measures during tanker shipping, and will be of significant managerial

contributions to the safety of tanker shipping industry.

ACKNOWLEDGMENT

This work was supported by the Singapore Maritime Institute under SMI Simulation & Modelling R&D Programme.

REFERENCES

[1] BP Statistical Review of World Energy, June 2015. Available: http://www.bp.com/content/dam/bp/pdf/Energy-economics/statistical-review-2015/bp-statistical-review-of-world-energy-2015-full-report.pdf

[2] Witten, Ian H., and Eibe Frank. Data Mining: Practical machine learning tools and techniques. 2nd ed. Morgan Kaufmann, 2005.

[3] Hastie, Trevor, et al. "The elements of statistical learning: data mining, inference and prediction." The Mathematical Intelligencer, vol.27 (2005):pp 83-85. http://dx.doi.org/10.1007/BF02985802

[4] Bevilacqua, M., F. E. Ciarapica, and G. Giacchetta. "Industrial and occupational ergonomics in the petrochemical process industry: A regression trees approach." Accident Analysis & Prevention, vol.40 (2008): pp1468-1479. http://dx.doi.org/10.1016/j.aap.2008.03.012

[5] Matías, J. M., et al. "A machine learning methodology for the analysis of workplace accidents." International Journal of Computer Mathematics 85.3-4 (2008): pp559-578.

[6] Nenonen, Noora. "Analysing factors related to slipping, stumbling, and falling accidents at work: Application of data mining methods to Finnish occupational accidents and diseases statistics database." Applied ergonomics, vol.44 (2013): pp215-224. http://dx.doi.org/10.1016/j.apergo.2012.07.001

[7] Cheng, Ching-Wu, et al. "Applying data mining techniques to explore factors contributing to occupational injuries in Taiwan's construction industry."Accident Analysis & Prevention, vol.48 (2012):pp 214-222. http://dx.doi.org/10.1016/j.aap.2011.04.014

[8] Bevilacqua, Maurizio, Filippo Emanuele Ciarapica, and Giancarlo Giacchetta. "Data mining for occupational injury risk: a case study." International Journal of Reliability, Quality and Safety Engineering, vol.17 (2010): pp351-380. http://dx.doi.org/10.1142/S021853931000386X

[9] Rivas, T., et al. "Explaining and predicting workplace accidents using data-mining techniques." Reliability Engineering & System Safety, vol.96 (2011): 739-747. http://dx.doi.org/10.1016/j.ress.2011.03.006

[10] Fragiadakis, N. G., V. D. Tsoukalas, and V. J. Papazoglou. "An adaptive neuro-fuzzy inference system (anfis) model for assessing occupational risk in the shipbuilding industry." Safety Science, vol.63 (2014): pp226-235. http://dx.doi.org/10.1016/j.ssci.2013.11.013

[11] Takagi, Tomohiro, and Michio Sugeno. "Derivation of fuzzy control rules from human operator's control actions." Proceedings of the IFAC symposium on fuzzy information, knowledge representation and decision analysis. Vol. 6. 1983.

[12] Witten IH, Frank E. Data mining: practical machine learning tools and techniques. 2nd ed. Morgan Kaufmann; 2005.

[13] Breiman L, Friedman J, Olshen R, Stone C. Classification and regression trees. Wadsworth; 1984.

[14] Ciarapica, F. E., and G. Giacchetta. "Classification and prediction of occupational injury risk using soft computing techniques: an Italian study." Safety science, vol.47 (2009): pp36-49. http://dx.doi.org/10.1016/j.ssci.2008.01.006

[15] Hofleitner, Aude, et al. "Learning the dynamics of arterial traffic from probe data using a dynamic Bayesian network." Intelligent Transportation Systems, IEEE Transactions on, vol.13 (2012): pp1679-1693. http://dx.doi.org/10.1109/TITS.2012.2200474

[16] Larasati, Aisyah, Camille DeYong, and Lisa Slevitch. "The application of neural network and logistics regression models on predicting customer satisfaction in a student-operated restaurant." Procedia-Social and Behavioral Sciences, vol.65 (2012):pp 94-99. http://dx.doi.org/10.1016/j.sbspro.2012.11.097

[17] Mezghani, Neila, et al. "Kinematic gait analysis of workers exposed to knee straining postures by Bayes decision rule." Artificial Intelligence Research, vol.4 (2015): pp106-111. http://dx.doi.org/10.5430/air.v4n2p106