Study of Optimized SVM for Incident Prediction of a Steel Plant in India

Conference Paper · December 2016

DOI: 10.1109/INDICON.2016.7838894

Sobhan Sarkar
Indian Institute of Technology Kharagpur
20 PUBLICATIONS **26** CITATIONS

SEE PROFILE

Sammangi Vinay
Indian Institute of Technology Kharagpur
3 PUBLICATIONS **9** CITATIONS

SEE PROFILE

Vishal Pateshwari
Indian Institute of Technology Kharagpur
2 PUBLICATIONS **3** CITATIONS

SEE PROFILE

Study of Optimized SVM for Incident Prediction of a Steel Plant in India

Sobhan Sarkar
Research Scholar
Department of Industrial and Systems Engineering,
Indian Institute of Technology Kharagpur,
sobhan.sarkar@gmail.com

Sammangi Vinay
B. Tech.
Department of Mechanical Engineering
Indian Institute of Technology,
Kharagpur, India.
sammangi.vinay@gmail.com

Vishal Pateshwari
M. Sc.
Department of Geology and Geophysics,
Indian Institute of Technology,
Kharagpur, India.
vishal.pateshwari@gmail.com

Jhareswar Maiti
Professor
Department of Industrial and Systems Engineering,
Indian Institute of Technology, Kharagpur
jhareswar.maiti@gmail.com

Abstract— **Occupational accident is a serious issue for every industry. Steel industry is considered to be one of the economic sectors having a high number of accidents. Thus, the main aim of this study is to build a model which could predict the occupational incidents (i.e., injury, near-miss, and property damage) using support vector machine (SVM) by utilizing a database comprising almost 5000 occupational accidents reports from an integrated steel plant corresponding to the span of years 2010 to 2012. Parameter optimization of the SVM is performed using grid search (GS), genetic algorithm (GA), and BAT algorithm to obtain the better accuracy of the classifier. The results of experiments show that grid search-based SVM outperforms other optimized SVM approaches with 88.0% accuracy. Other optimization techniques can also be adapted to search for the better prediction accuracy of the model.**

Keywords - **Occupational accident, Classification, SVM, Parameter optimization, Grid search, Genetic algorithm, BAT algorithm.**

I. INTRODUCTION

Occupational accidents occur unexpectedly and in a sudden manner. These are occurred due to any violence, death, or any accidents in a workplace resulting in the personal injury of workers at the site. In an average, estimated fatal occupational accidents was 14.0 per 100,000 workers, and the estimated number of those accidents was 335,000 world-wide in 1994 [1]. According to International Labor Organization [2], it is estimated that 2.3 million lose their life due to work related diseases and accidents, in which around 360,000 die due to fatal accidents. Overall, every year approximately 337 million accidents occur at work [2]. Due to these accidents, there is nearly 4% loss in global Gross Domestic Product (GDP), which is nearly equal to US $1.25 trillion, due to the interruption in production, medical coverage, labor compensation and loss in working time [2]. According to EUROSTAT, more than 5700 people lose their life due to occupational accidents every year in Europe [3]. In European Union (EU)-15 in 2004, it was estimated that, 55 billion euros each year were spent for accidents and in EU-27, 3.2% workers which is equal to 7 million workers encounter an accident at work [4]. Thus, it is evident that occupational accidents create a serious economic burden to the industry as well as the country.

To sort out these problems, many researchers use different conventional techniques like logistic regression [5], Bayesian network [6], etc. With the advancement of technology, many data-driven learning based models using machine learning (ML) techniques have been developed. ML is being widely used in several domains like information retrieval, image processing, business, health care etc. These approaches are very effective in data analysis. The algorithm learns itself from the data and allows the computer to find hidden insights without any type of explicit programming. These techniques are found out to be very potential while dealing with huge amount of data as they can provide quick and more accurate output than other approaches. In accident prediction, many researches have been carried out using the state-of-the art approaches in line with ML. For instances, fault tree based Bayesian networks [7], decision tree [8], classification approaches using k-nearest neighbor (k-NN) [9], artificial neural network (ANN) [10], support vector machines (SVM) [11], are used frequently in prediction of accidents. Out of all the mentioned approaches, SVM is one of the most commonly used techniques for classification purpose as it can perform nonlinear classification efficiently [12]. Though, SVM works well with considerable accuracy depending on dataset, but it alone cannot efficiently recognize new data [12]. Previously SVM has been used in road accident analysis [13], landing safety analysis [14] and safety analysis of human accidents [15]. Parameter optimization is a technique that employs the search of optimal values of the parameters from the solution space resulting in improved performance measure of the learning algorithm. To tackle this issue, the main purpose of the present work is set to build a predictive model for occupational accidents using SVM optimized by grid search (GS), genetic algorithm (GA) and BAT algorithm (BAT) techniques in order to search for higher prediction accuracy. The scope of the paper is limited only to experiment

978-1-5090-3646-2/16/$31.00 ©2016 IEEE.

on the SVM with optimization algorithms (GS, GA, and BAT) applied on its parameters towards the prediction of occupational incidents. However, selection of optimization algorithms is based on the previous studies in different fields. SVM is found to be frequently used algorithm in classification problem, either in accident prediction [9], or image classification, or regression problems with small samples having large dimensional features [16]. It is also found from the literature that optimal parameter selection could enhance the accuracy of the classifiers [16-17]. In fact, selection of suitable parameters is usually treated as an optimization problem. Thus, to achieve the best accuracy from SVM, parameter optimization is required. In literature, many evolutionary algorithms have been used with SVM to optimize the parameters in such a way so that accuracy of the classifier should get increased. Many evolutionary algorithms like GA [17], particle swarm optimization (PSO) [18] have been used frequently in different types of problems and their results are also found out to be better depending on the problem at hand. Another new heuristic algorithm called BAT algorithm, which allows stochastic search and holds very good performance while dealing with optimization problems [19]. Sometimes, BAT algorithm outperforms GA, and PSO algorithm also [16]. Therefore, in the present study, to obtain the better accuracy in classification problem, grid search, genetic algorithm, and BAT algorithm have been chosen for experimentation purpose. Finally, the results of the grid search optimized SVM (GS-SVM), genetic algorithm optimized SVM (GA-SVM) and BAT algorithm optimized SVM (BAT-SVM) are compared. Hence, experimentation results are promising, and hold a potential application of our algorithm for prediction problem of the occupational accidents.

The rest of the paper is organized as follows: Section II contains the contribution of work. The problem statement of the case study is described in Section III. The methodological flowchart and methods used in this paper are discussed in Section IV. Results and discussion is provided in Section V, and finally, Section VI provides the conclusion.

II. CONTRIBUTION OF THE WORK

Due to the space limitation, some of the important previous works on occupational accident prediction are illustrated in previous section. Major focus of previous works on prediction of occupational accidents has been primarily limited only to the implementation of single machine learning algorithm towards the problems at hand. But, optimization of parameters corresponding to the algorithm has hardly been reported in earlier works. Taking the factor into consideration, this study proposes an optimized SVM method, which parameters are optimized by GS, GA and BAT techniques. Hence, in the field of occupational accident prediction, to the best of the authors' insight, the proposed approach is new and holds a good potential for obtaining an improved accuracy than using single learner like SVM.

III. PROBLEM STATEMENT

The study made here is focused primarily on providing a predictive solution to an integrated steel plant (in India) having problem related to occupational accidents. The company has the records of accidents taken place in the plant in the years spanning from 2010 to 2012. But, due to the lack of efficient analysis of the recorded data, the accidents, within the workplace, could not be controlled, which, in turn leads to create an unsafe environment resulting in injuries, damage of property, or near-miss cases. So, to reduce the occurrence of accidents in the industry, the safety management system opts to use the predictive analytics which is believed to be a powerful approach towards the accident prevention strategy making.

A. Dataset

The dataset contains the details of events of three years (2010 –2012). It has 31 attributes initially. In data pre-processing stage, only 16 important attributes are chosen for analysis using experts' judgments. All the attributes and their nature are mentioned below:

Division (categorical), Department (categorical), Injury type (categorical), Primary cause (categorical), Status (categorical), Working condition (categorical), Machine condition (categorical), Observation type (categorical), Employee type (categorical), Serious process incident score (numerical), Injury potential score (numerical), Equipment damage score (numerical), Safety standards (categorical), Incident type (categorical), Standard Operation Procedure (SOP) (categorical), Incident category (categorical). All the numerical attributes are converted into categorical on the basis of expert's judgments.

IV. METHODS USED

The proposed methodological flowchart is depicted in Fig. 1. The methods used in this paper are briefly discussed below.

A. Data preprocessing

Data preprocessing is an essential part of data mining. Quality of results is directly proportional to quality of data. Missing data (nearly 40% of total data set) was filled up by contextual information mapping. Data transformation was applied to numeric attributes in order to convert them into categorical attributes by experts' judgment for improving the accuracy of the analysis.

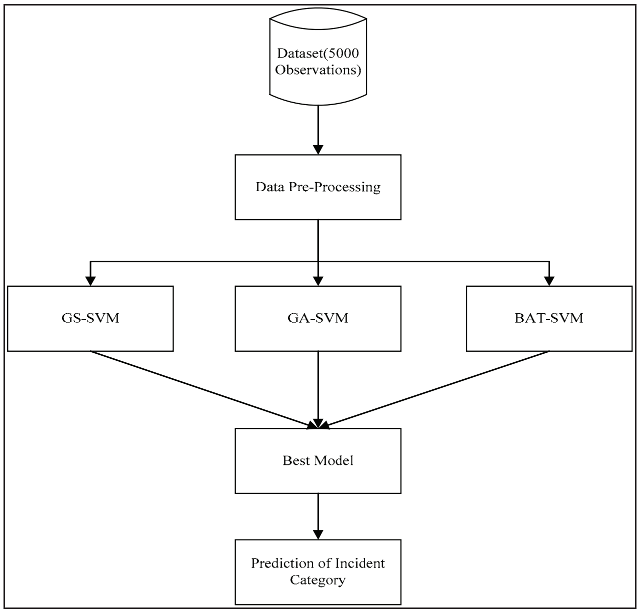

Fig. 1. Methodological flow chart.

B. *Support vector machine (SVM)*

SVM is one of the most important classification techniques. It is designed on the basis of Vapnik-Chervonenkis Theory [20-21]. It classifies the data having different class-labels by finding a set of support vectors that are set members of inputs of training set. SVM has certain parameters (i.e., cost and gamma) which are to be simulated for the best result. Cost and gamma are kernel based parameters. Initially, the training dataset has been used for training the prediction model and the testing dataset has been used to feed the prediction model for the calculation of accuracy.

Finding out the linear hyper-planes to separate the data in different classes with the best possible margin is the main objective of SVM. Different class data is separated in such a way that the margin separating them has the maximal distance with the nearest data points and the hyper-plane is bounded by the margin. Support Vectors are the points that are closest to the hyper-plane. These data points are important as they consist of all the data require to model the classifier.

The hyper-plane is written as $a_1^T x_i + a_2 = 0$, and then conversion of SVM classification problem is compassed to find out the optimal parameters; a_1^T and a_2 to maximize the distance between hyper-planes $a_1^T x_i + a_2 = 1$ and $a_1^T x_i + a_2 = -1$. In accordance with [17], the equation transforms to following optimization problem:

$$\min_{a_1,a_2} \frac{1}{2} || a_1 ||^2 + C\sum_{i=1}^{n} \xi_i \qquad (1)$$

$$\text{s.t. } y_i(a_1^T x_i + a_2) \geq 1 - \xi_i \qquad (2)$$

$$\xi_i \geq 0, \qquad (3)$$

In Eq. (1), parameter C is called cost and it is a user-specified parameter for error term. The parameter ξ_i represents the error term. There are cases in which the data cannot be separated linearly. Kernel function is used in such cases. Linear kernel and radial basis function are some of the commonly used kernel functions. Gaussian radial basis function has been used in this paper as it can perform alike with linear kernel as well as sigmoid kernel and also that the RBF kernel considers only cost(C), and kernel parameter gamma (γ).

The aim of performing the cross validation is to characterize a dataset in order to test the model in the training stage, to overcome the problems such as over fitting. This provides an insight to generalize the model into an independent dataset. The RBF kernel on two samples ‘a’ and ‘b’, described as the feature vectors in any input space, is construed as:

$$K(a,b) = e^{\left(-\frac{||a-b||^2}{2\times\sigma^2}\right)} \qquad (4)$$

where $||a\text{-}b||^2$ is the squared Eucledian distance between two feature vectors and σ is a free parameter. SVM with grid search has been used in order to tune the parameters i.e., gamma and cost.

C. *Grid Search (GS)*

Grid Search (GS) is a simple, yet one of the best methods to find optimal parameter values for SVM model. Cost (C) and gamma (γ) are two independent parameters of SVM. A range for C and γ is set first then the values from the range specified are taken as a pair and tested one by one and is assessed by cross-validation. The pair with the best accuracy or pair with minimum error rate is returned as the optimal parameter values. According to previous works [22], C and γ are selected in an exponentially increasing way.

D. *Genetic Algorithm (GA)*

Genetic Algorithm (GA) is an algorithm that mimics the process of natural selection in the field of mathematical optimization. It is generally used to generate useful results to optimization problems. In a problem solving, GA initially encodes the potential solutions into chromosomes. Then by the combination of all chromosomes that are created, a search space is created. Then in next step, random selection of chromosomes takes place to form initial population. Now, a fitness value is assigned to each chromosome by calling it as the fitness function. The chromosomes which are having highest fitness value are signified as the better ones. In the next stage, to give birth to new generation of chromosomes GA operates selection, mutation and crossover. The fitness values that are expected are supposed to be higher than the previous generation. When the termination criterion is satisfied, GA stops and the chromosome with highest fitness in latest generation is considered as the best possible solution.

E. Bat Algorithm

The approach of this algorithm is to mimic the performance of bats when catching their prey. This was influenced by the echolocation nature of the bats. BAT was first conferred in [23]. In order to outperform Genetic Algorithm and Particle Swarm Optimization in assessment, it was developed by using some of the benchmark functions. The pulse rates of loudness and emission are varied. It is also used to solve hard optimization problems like global engineering optimization [24], clustering complications [25], and constrained optimization tasks [26]. For feature selection, two versions of bat-influenced algorithms were proposed [27]. The application of this algorithm is more difficult than numerous meta-heuristic algorithms as each bat is assigned a set of combining parameters i.e., frequencies, loudness, position, rate of pulse emission and velocity. These combinations alter the nature as well as the time required to achieve the solution.

The fundamental principle of the algorithm is discussed here. A group of bats are assumed to randomly fly at position x_i, a fixed frequency f, velocity v_i, loudness a_0 and changing wavelength λ for searching the prey. They have the capacity to alter the wavelength of the pulses that are emitted and also to adjust the pulse rate, $r_p \in [0,1]$ which is crucial to find their proximity of the target. The frequency varies from $f_{\min}$ to $f_{\max}$ which means that the wave lengths vary from $\lambda_{\min}$ to $\lambda_{\max}$. Even though the loudness is contrast in different ways, it alters from a_0 which is large and also positive to a minimum and constant value i.e., $a_{\min}$. This algorithm is widely used for global optimization.

V. RESULTS AND DISCUSSION

Here, in this section, few key findings in our study have been provided in details. Both the categorical and numerical attributes from the data are considered for analysis. In the data pre-processing stage, the numerical attributes were converted into categorical attributes. Then multi-class SVM classifier using 10-fold cross validation (CV) with different parameter optimization techniques are implemented onto the data to predict the incident category. In this experiment, "ranges = list (gamma = 2^{-5} *to* 2^{5}, cost = 2^{-5} *to* 2^{5})" are kept in tune function so that we get accuracies for 121 posssible cases. Fig. 2 shows the variation of accuracies with cost being fixed and gamma being varied from 2^{-5} *to* 2^{5}. Similarly, Fig.3 shows the accuracies with gamma being fixed and cost being varied from 2^{-5} *to* 2^{5}.

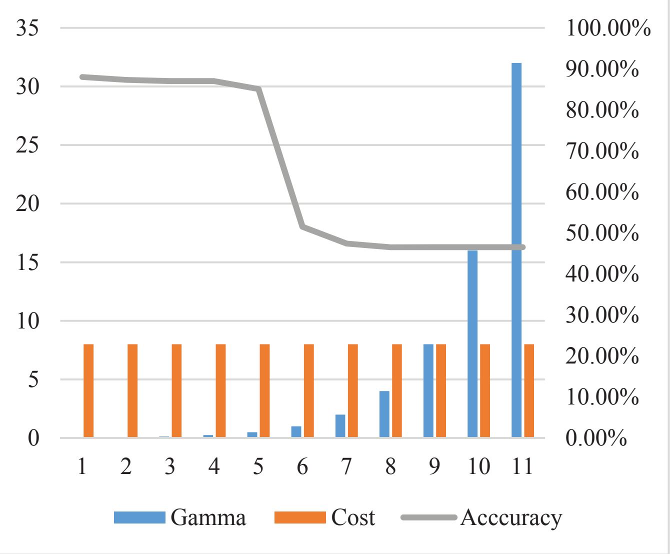

Fig. 2. Variation of accuracies with gamma in GS-SVM.

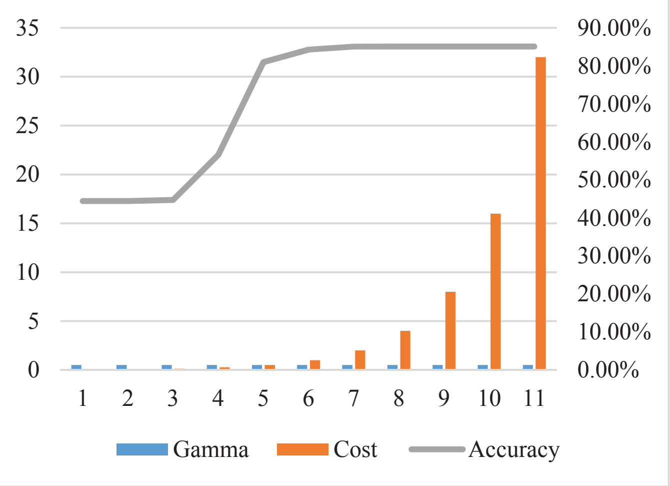

Fig. 3. Variation of accuracies with cost in GS-SVM.

From Fig. 2, we can see that the accuracies are decreasing with gamma when cost is fixed. Similarly, from Fig. 3, we can also see that accuracies are increasing with cost when gamma is fixed. Out of all 121 possible cases, the best accuarcy is found out to be 88.0% for which gamma=0.03125 and cost=8.

Also, SVM with GA has been used for analyzing the data. In GA, fitness function is used as the SVM prediction model, which returns the accuracy of that model, with a population size of 25 and maxinum number of iterations being 10. The chromosome with the best fitness value (accuracy) is returned as the optimized result after selection, mutation and crossover. The Fitness value vs Generation (iteration) graph is shown in Fig. 4.

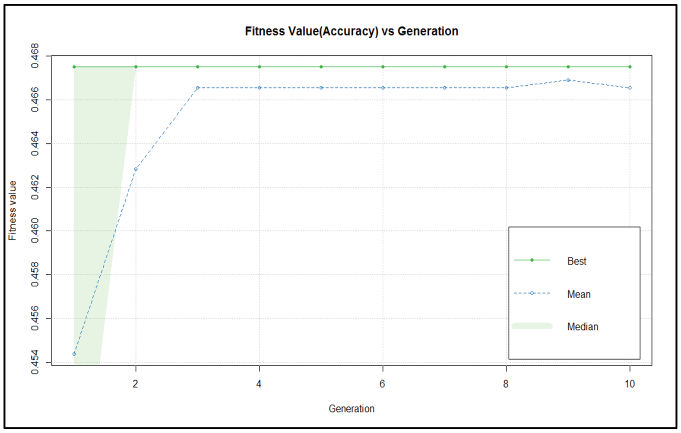

Fig. 4. Fitness vs. generation of GA-SVM model.

The best accuracy or the fitness value turns out to be as low as 46.75% in this method.

BAT algorithm has also been implemented on multiclass SVM classifier. It returns the accuracy as the fitness function. The performace is calculated for the parameter settings shown in TABLE I. The best accuracy is found to be 87.2%.

TABLE I: PARAMETER SETTINGS USED FOR IMPLEMENTATION OF BAT ALGORITHM.

SN	Parameter	Value
1	Dimension of search variables, d	10
2	Population size, n_p	25
3	Number of generations, n_{gen}	10
4	Minimum Frequency, f_{min}	0
5	Minimum Frequency, f_{max}	2
6	Loudness, a	0.5
7	Lower bound of search variables	-10
8	Upper bound of search variables	10
9	Rate of pulse emission, r_p	0.5
10	Objective function, fun	accuracy

TABLE II shows the comparison of accuracies of all the parameter optimization techniqueds used. It is seen that GS-SVM predicts the incident category better than other two methods used.

Just like other research, this study too has some limitations. One of them is that the algorithm approach has not been used for handling the missing values. Second, feature selection has been done manually which can be done by algorithmic approach. Moreover, in GA, the maximum iteration has been considered as 10 with population size 25 which could have also been set as higher as possible for better visualization the results of simulation.

TABLE II: COMPARISON OF ACCURACIES OF SVM AND OPTIMISATION TECHNIQUES OF SVM.

SN	Method (10 fold CV)	Accuracy
1	GS-SVM	88.00%
2	GA-SVM	46.75%
3	BAT-SVM	87.20%

V. CONCLUSION

This study discussed the SVM classifier and its parameter optimization techniques. It can be noticed that the performance of GS-based SVM model is higher than the other two methods used. It is seen that the accuracy of GS-SVM is 88.0%, which is higher than others. So, GS-SVM can be used for predicting the incident category in steel industries. For the future work, other optimization techniques such as ant colony optimization (ACO) and PSO can also be applied on SVM classifier to check the prediction capability of the model. The other methods like ANN, k-NN, multi-layer perceptron (MLP), decision trees could also be tried with proper parameter optimization techniques to achieve better prediction accuracy. In GA, population size and iteration could be taken as much as possible considering time and cost complexity parameters. Moreover, the data points could also be collected as much as possible for obtaining better prediction accuracy.

ACKNOWLEDGMENT

The authors are thankful to the safety management of the steel industry for their support in this research.

REFERENCES

[1] J. Takala, "Global estimates of fatal occupational accidents. *Epidemiology-Baltimore"*, vol. *10, no.* 5, pp. 640-646, 1999.

[2] International Labour Organization (ILO), Promoting safe and healthy jobs, The ILO Global Programme on Safety, Health and the Environment (Safework), in: World of Work, vol. 63, 2008, pp. 4–11.

[3] EUROSTAT, Labour Force Survey 2007 ad hoc module on accidents at work and work-related health problems, in: European Communities, 2009

[4] European Commission, Directorate-General for Employment, Social Affairs and Equal Opportunities, The Social Situation in the European Union 2007, in: European Communities, 2007

[5] C. L. A. Taibo, T. D. Moon, O. A Joaquim, C. R. Machado, A. Merchant, K. McQueen & E. Folgosa, "Analysis of trauma admission data at an urban hospital in Maputo, Mozambique". *International journal of emergency medicine*, vol. *9, no.*1, pp. 1, 2016.

[6] E. Castillo, Z. Grande & A. Calviño, "Bayesian Networks-Based Probabilistic Safety Analysis for Railway Lines." *Computer-Aided Civil and Infrastructure Engineering*, 2016.

[7] S. Sarkar, S. Vinay, & J. Maiti, "Text mining based safety risk assessment and prediction of occupational accidents in a steel plant." In *Proceedings of the IEEE International Conference on Computational Techniques in Information and Communication Technologies (ICCTICT) 2016*, pp. 439-444, 2016.

[8] L. Sanmiquel, J. M. Rossell & C. Vintró, "Study of Spanish mining accidents using data mining techniques". *Safety science*, vol. *75*, pp. 49-55, 2015.

[9] Z. Chen, X. C. Liu, & G. Zhang, "Non-recurrent congestion analysis using data-driven spatiotemporal approach for information construction." *Transportation Research Part C: Emerging Technologies*, vol. *71*, pp. 19-31, 2016.

[10] W. Yi, A. P. Chan, X. Wang, & J. Wang, "Development of an early-warning system for site work in hot and humid environments: A case study." *Automation in Construction*, vol. *62*, pp. 101-113, 2016.

[11] A. S. Sánchez, P. R Fernández, F. S. Lasheras, F. J. de Cos Juez, & P. G. Nieto, "Prediction of work-related accidents according to working conditions using support vector machines". *Applied Mathematics and Computation*, vol. *218, no.* 7, pp. 3539-3552, 2011.

[12] B. M. Aslahi-Shahri, R. Rahmani, M. Chizari, A. Maralani, M. Eslami, M. J. Golkar, & A. Ebrahimi, "A hybrid method consisting of GA and SVM for intrusion detection system." *Neural Computing and Applications*, pp. 1-8, 2015.

[13] B. Sharma, V. K. Katiyar, & K. Kumar, "Traffic Accident Prediction Model Using Support Vector Machines with Gaussian Kernel." In *Proceedings of Fifth International Conference on Soft Computing for Problem Solving* (pp. 1-10). 2016, Springer Singapore.

[14] Y. Dai, J. Tian, H. Rong, & T. Zhao, "Hybrid safety analysis method based on SVM and RST: An application to carrier landing of aircraft." *Safety science*, vol. *80*, pp. 56-65, 2015.

[15] J. Zhu, & M. Xiao-ping, "Safety evaluation of human accidents in coal mine based on ant colony optimization and SVM." *Procedia Earth and Planetary Science*, vol. *1*, no.1, pp. 1418-1424, 2009.

[16] Z. Ye, L. Ma, M. Wang, H. Chen, & W. Zhao. "Texture image classification based on support vector machine and bat algorithm." In *Intelligent Data Acquisition and Advanced Computing Systems: Technology and Applications (IDAACS), 2015 IEEE 8th International Conference on* (Vol. 1, pp. 309-314). IEEE.

[17] Q. Zhang, G. Shan, X. Duan, Z. Zhang, "Parameters optimization of support vector machine based on simulated annealing and genetic algorithm," in Proceedings of the IEEE International Conference on Future Computer and Communication, 2009, pp. 282-284.

[18] S. Zhai, T. Jiang, "A novel particle swarm optimization trained support vector machine for automatic sense-through-foliage target recognition system," Knowledge-Based Systems, vol. 65, 2014, pp. 50-59.

[19] X. S. Yang, "A new metaheuristic bat-inspired algorithm," in Nature Inspired Cooperative Strategies for Optimization (NISCO'2010) (Eds. J. R. Gonzalez et al.), Studies in 314 Computational Intelligence, Springer, Berlin, vol. 284, 2010, pp. 65-74.

[20] B.E. Boser, I.M. Guyon, & V.N. Vapnik, "A training algorithm for optimal margin classifiers", In: Proceedings of the fifth annual workshop on computational learning theory, ACM, New York, 1992, pp. 144-152.

[21] S.R. Gunn, Support vector machine for classification and regression, ISIS Technical Report, 14, 1998.

[22] C. W. Hsu, C. C. Chang, & C. J. Lin, A practical guide to support vector classification, 2003.

[23] C. Cruz, J. R. González, D. A. Pelta, N. Krasnogor, & G. Terrazas, (Eds.) "*Nature Inspired Cooperative Strategies for Optimization" (NICSO 2010)* (Vol. 284). 2010. Springer.

[24] X. S. Yang, & A. Hossein Gandomi, "Bat algorithm: a novel approach for global engineering optimization." *Engineering Computations*, vol. *29, no.* 5, pp.464-483, 2012.

[25] K. Khan, A. Nikov, & A. Sahai, "A fuzzy bat clustering method for ergonomic screening of office workplaces." In *Third International Conference on Software, Services and Semantic Technologies S3T 2011* (pp. 59-66). Springer Berlin Heidelberg.

[26] A. H. Gandomi, X. S. Yang, A. H. Alavi, & S. Talatahari, "Bat algorithm for constrained optimization tasks." *Neural Computing and Applications*, vol. *22, no.* 6, pp. 1239-1255, 2013.

[27] A. M. Taha, & A. Y. Tang, "Bat algorithm for rough set attribute reduction." *Journal of Theoretical and Applied Information Technology*, vol. 51, no. 1, pp. 1-8, 2013.