Available online at www.sciencedirect.com

ScienceDirect

Procedia Engineering 71 (2014) 328 – 332

Procedia Engineering

www.elsevier.com/locate/procedia

Data Mining on Fire Records of New South Wales, Sydney

Eric Wai-ming Lee[a,*], Guan-heng Yeoh[b], Morgan Cook[c], Chris Lewis[c]

[a] *Deparrment of Civil and Architectural Engineering, City University of Hong Kong, Hong Kong, China*
[b] *School of Mechanical and Manufacturing Engineering, University of New South Wales, Kensington, NSW 2052, Australia*
[c] *Fire & Rescue NSW, Greenacre, NSW 2190, Australia*

Abstract

This study gathered fire records from the Fire and Rescue New South Wales (F&RNSW) for investigating the most relevant event to the fire accident. Support vector machine was adopted to mimic the correlation between the information of the building and occupants and the occurrence of fire accident. The percentage of correct prediction is 65% which is considered reasonable since noise is expected to be embedded in the data of the fire records. Bayesian approach was also adopted to analyze the relevancies of the binary input parameters to the fire occurrence. Monte Carlo simulation was conducted. The result shows that the Special-Risk-Building and Smokers are the two parameters most relevant to the occurrence of fire accident.

Peer-review under responsibility of School of Engineering of Sun Yat-Sun University

Keywords: bayesian theorem, fire records, monte carlo simulation, support vector machine.

Nomenclature

A	fire accident
$\boldsymbol{b}$	constant column matri
F	either one of the binary inputs of the fire records
$\boldsymbol{r}$	vector from point $\boldsymbol{x}'$ to point $\boldsymbol{x}$
$\boldsymbol{w}$	weight matrix
$\boldsymbol{x}$	sample
$\boldsymbol{x}'$	point at the decision boundary nearest to the sample $\boldsymbol{x}$
y	label of the sample class
Greek symbols	
α_i	thei^{th} Lagrange multiplier
δ	time angle
θ	date angle
ρ	total width of the margins

1. Introduction

Most of the current fire investigations on the causes of fires rely on the practical experiences of the fire investigators. They make their judgment according to the background information of the buildings, the occupants and the evidences left in the fire scenes. Different fire investigators may arrive different conclusions due to their different experiences in fire investigations. This paper proposes to apply support vector machine (SVM) to objectively determine whether a fire is an

* Corresponding author. Tel.: +852-3442-2307; fax: +852-3442-0427.
E-mail address: ericlee@cityu.edu.hk

Peer-review under responsibility of School of Engineering of Sun Yat-Sun University
doi:10.1016/j.proeng.2014.04.047

accident or not. Also, Bayesian approach was adopted to evaluate the relevancies of different parameters to the fire accident from which precaution measures can be developed by the fire bridges to prevent the occurrences of the fire accidents. SVM was firstly developed by Vapnik[1]. It is a statistical learning model for classification task. It draws a decision boundary in the domain of the data to demarcate different groups of data by maximizing the margins between the decision boundary and the data points. The data points being used to establish the decision boundary are called support vectors. The SVM has been proven to be robust [2] in model training. Its performance was found to be superior to the traditional artificial neural network models (e.g. multilayer perceptron, radial basis function, general regression neural network, etc.).

Fire and Rescue New South Wales (F&RNSW) is the fire department in New South Wales (NSW), Sydney, Australia. They have a comprehensive system to record the fire cases occurred in NSW. This research obtained the fire records from their system to train a SVM model fordetermining the nature of a fire (i.e. accident or not) without any human intervention.

2. Brief Review on Support Vector Machine (SVM)

The SVM is a statistical learning model for classification. In the case of linear decision boundary, it draws the boundary on the sample domain to separate samples from two different classes. Assume the linear decision boundary is $\boldsymbol{w}^T\boldsymbol{x} + \boldsymbol{b} = 0$. As shown in Fig. 1 where $\boldsymbol{x}'$ is the nearest point at the decision boundary, the shortest distance between a sample x and the decision boundary is $\boldsymbol{r} = \mathrm{r}\boldsymbol{w}/|\boldsymbol{w}|$. The distance is evaluated as $\mathrm{r} = (\boldsymbol{w}^{\mathrm{T}}\boldsymbol{x} + \boldsymbol{b})/|\boldsymbol{w}|$. If we denote the points inside and outside the margin to be $y = -1$ and $y = +1$ respectively, we have the following equation.

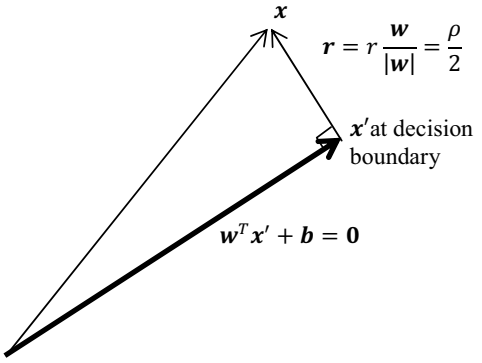

Fig. 1.Point x is a sample and $\boldsymbol{w}^T\boldsymbol{x}' + \boldsymbol{b} = 0$ is the linear decision boundary. The shortest distance from the sample to the decision boundary is denoted as margin which equals to $\rho/2$. The determination of the decision boundary is to maximize the margin.

$$\begin{cases} \boldsymbol{w}^T\boldsymbol{x} + \boldsymbol{b} \le -\dfrac{\rho}{2} & if\ y = -1 \\ \boldsymbol{w}^T\boldsymbol{x} + \boldsymbol{b} \ge \dfrac{\rho}{2} & if\ y = +1 \end{cases}$$

By combining the above two equations, we have $y(\boldsymbol{w}^T\boldsymbol{x} + \boldsymbol{b}) \ge \rho/2$. For the samples form the margins (i.e. support vectors), the equation becomes $y(\boldsymbol{w}^T\boldsymbol{x} + \boldsymbol{b}) = \rho/2$. The equation is scaled by $\rho/2$ and it becomes $\boldsymbol{w}^T\boldsymbol{x} + \boldsymbol{b} = 1$. Therefore, we have the margin $r = 1/|\boldsymbol{w}|$and $\rho = 2r = 2/|\boldsymbol{w}|$. It shows that maximizing the margin is equivalent to minimizing the value of $|\boldsymbol{w}|$or $\boldsymbol{w}^T\boldsymbol{w}/2$. The dual problem is solved by Lagrange multiplier as follows.

$$Max\left(\sum_i \alpha_i - \frac{1}{2}\sum_i\sum_j \alpha_i\alpha_j y_i y_j\, \boldsymbol{x}_i^T\boldsymbol{x}_j\right)\text{with the following conditions.}$$

$$\begin{cases} \sum_i \alpha_i y_i = 0 & Condition\ 1 \\ \alpha_i \ge 0\ \text{f oral l}\alpha_i & Condition\ 2 \end{cases}$$

It results the following equations for the determination of the values of α_j.

$$\begin{cases} 1 - y_i\sum_j \alpha_j y_j\,\langle \boldsymbol{x}_i, \boldsymbol{x}_j\rangle = 0 \\ y_i\sum_j \alpha_j y_j\,\langle \boldsymbol{x}_i, \boldsymbol{x}_j\rangle = 1 \end{cases}$$

3. Fire Records and Data Preprocessing

The fire records were provided by F&RNSW. It summarized the fire cases occurred from year 2000 to 2011. Each fire record documented the background information of the building, the occupants and also the result of investigation. These records are important to the fire investigator to formulate future rules or strategies to prevent fire occurrence. Currently, the interpretation from the fire records was conducted according to the experiences of the fire investigation. Recently, the database becomes bigger and bigger. Solely relying on the fire investigator to interpret the data according to their experiences becomes unrealistic. An objective tool should be developed to provide advice to the fire investigation in their decision making. The selected major parameters contributing to the cause of fire in this study are summarized in Table 1.

Table 1. Summary of the input and output parameters adopted in this study

No.	Item	Value	Description
1	Visiting	Yes/No	Is there any visitor in the occupants?
2	Disabled	Yes/No	Is there any disabled person in the occupants?
3	Smoker	Yes/No	Is there any smoker in the occupants?
4	Alcohol	Yes/No	Is there any drinker in the occupants?
5	Drugs	Yes/No	Is there any occupant taking drug?
6	Medicines	Yes/No	Is there any occupant taking medicine?
7	Mental illness	Yes/No	Is there any occupant with mental illness?
8	Hoarder	Yes/No	Is there any occupant a hoarder?
9	Special risk	Yes/No	Is the building a specified special risk building?
10	Time	hh:mm	The time of the fire occurrence
11	Date	DD/MM/YY	The date of the fire occurrence
12	Accident	Yes/No	Is the fire an accident?

Referring to Table 1, the items 1 to 11 were taken to be the inputs of the SVM model. The output of the model is item 12. That is, the SVM model will determine whether the fire is accident or not based on the information provided in item 1 to 11. There were total 317 samples available to this study.Apart from item 10 and 11, the other parameters were binary data which can easily be represented by 1 (i.e. yes) or 0 (i.e. No). However, the date and time of the fire cases should be further processed before the actual application. Solely inputting the date, month, year, hour and minute into the SVM model may confuse the model since it is unable to describe that the day 1 January is just one day after 31 December, and, the time 23:59 is just 1 minute before 00:00. In order to do this, data pre-processing using cyclic concept was developed. We proposed to map the dates from the 1st January of one year to the 1st January of next year to a circle from 0° to 360°. Therefore, any date within a year could be mapped to an angle θ between 0° and 360°. A dual presentation as $\left(sin(\theta), cos(\theta)\right)$ is therefore sufficient to describe the unique date within a year. Similar technique was applied to convert the time to the dual presentation. We mapped the time from 00:00 to 24:00 to an angle δbetween 0° and 360°. A duplex presentation as $\left(sin(\delta), cos(\delta)\right)$ could describe the unique time within a day.

4. SVM Model Training and Results

In this study, the kernels of the SVM model were assumed to be Gaussian radial basis function. The support vectors were determined by the model itself while the spreads of the kernels were obtained by quadratic programming. Due to the limited number of samples (i.e. 317 fire cases), leave-one-out validation approach was adopted. The procedures are described as follows. In this first trial, the 1st sample was taken out from the pool of samples as a test sample. The other 316 samples were used as training samples to train the SVM model. Upon the completion of model training, the trained SVM model was applied to predict the output of the test sample. In the 2nd trial, the 2nd sample was taken as test sample and the procedures repeated until the last sample was used as test sample. Therefore, total 317 predicted outputs were created from the 317 samples through the leave-one-out validation. The performance indicator of the SVM model, percentage of correct prediction, is obtained from dividing the number of correct predictions by the total number of predictions (i.e. 317). Table 2 is the confusion matrix of the results obtained from the SVM model training and validation. It shows that, the actual accident fire cases were correctly predicted with 63.6% accuracy while the actual non-accident fire was correctly predicted with 66.4% accuracy. The two percentages are quite close to each other. The model did not bias to either one side of the classification problem. The overall percentage of correct prediction is 65% (i.e. 33% + 32%). The prediction error may due to the inaccuracy of the fire records since the records were taken by different fire investigators. They might have different

interpretations on the information they received and also their decisions as well. Therefore, noise is expected to be embedded in the collected real data. Therefore, the overall performance of the SVM model is considered reasonable.

Table 2. The confusion matrix summarizes the SVM prediction results

		PREDICTED		% of Correct Prediction
		Accident fire	Non-accident fire	
ACTUAL	Accident fire	105 (33%)	60 (19%)	$\frac{105}{105+6\ 0} = 63.6\%$
	Non-accident fire	51 (16%)	101 (32%)	$\frac{101}{51+101} = 66.4\%$

5. Discussion

The relevancy of each parameter to the fire accident was also investigated by Bayesian approach with Monte Carlo simulation. The relevancy is defined as how likely a parameter occurs if a fire is an accident. It can be described mathematically as $P(F|A)$ where F is the parameter and A is the fire accident. By Bayes theorem, it can be expressed as follows.

$$P(F|A) = \frac{P(A|F)P(F)}{P(A)}$$

The $P(F)$ is the probability of occurrence of parameter F in all fire cases. It can be obtained by the ratio of the number of fire cases with F to the total number of fire cases. The $P(A)$ is considered as a normalization factor. It can be ignored since we only concern on the relative relevancies between the parameters. Theoretically, the value of $P(A|F)$ can also obtained from the fire records by extracting all samples with $F = 1$ and counting the percentage of accident fire cases of the extracted samples (i.e. $A = 1$). However, it only represents the information in the sample space. We would propose to obtain the $P(A|F)$ by using the SVM model trained in section **错误！未找到引用源。** which may represent the property more close to the population space. Monte Carlo approach is adopted and the procedures are detailed as follows. When a parameter F (i.e. either one of the binary inputs in Table 1) is to be investigated, we create one million input vectors with $F = 1$ and with other input parameters randomly assigned. The one million input vectors are fed into the trained SVM model to generate one million predicted results. The percentage of predicted fire accident within the one million results therefore represents the value of $P(A|F)$. The above procedures are repeated for each of the 9 binary inputs as shown in Table 1. The results are summarized in Table 3.

Table 3. The confusion matrix summarizes the SVM prediction results

Input parameter, F	$P(A\|F)$	$P(F)$	$P(F\|A)P(A)$
Visiting	0.0095	0.9982	0.0094
Disabled	0.1136	0.9924	0.1127
Smoker	0.2839	0.9755	0.2770
Alcohol	0.2114	0.9794	0.2070
Drugs	0.0757	0.9838	0.0745
Medicines	0.1293	0.9875	0.1277
Mental illness	0.1104	0.9814	0.1084
Hoarder	0.0789	0.9925	0.0783
Special risk	0.3060	0.9812	0.3002

For easy visualization, all values in the column of$P(F|A)P(A)$ in Table 3 are normalized andgraphically presented in Fig. 2 to depict the relevant relevancies between the binary input parameters to occurrence of fire accident. It shows that the highest relevancy parameter is the Special-risk-building. A building is labelled special-risk-building by the NSW Government according to the usage of the building. This result indicates that the NSW Government may require to further enhance the existing fire protection measures of the special risk buildings. The second highest relevancy is Smoker. More

education to the occupants is required to reveal the seriousness of this issue to the community. Strategies through regulatory approach may also be considered. The information indicated in Fig. 2 is very useful to the NSW Government for their reference in formulating strategy to reduce the fire accident.

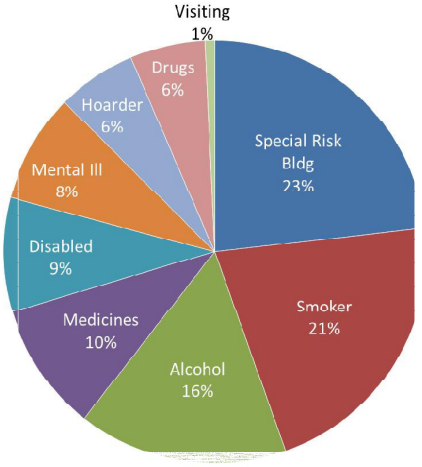

Fig. 2. The relative relevancies of different binary input parameters show that Special-risk-building and Smoker are the two major parameters relative more relevant to the occurrence of fire accident

6. Conclusions

This paper investigated the fire records provided by the F&RNSW. SVM was adopted to develop an intelligent model to predict the fire accident based on the parameters specified. The percentage of correct prediction achieves 65% which is considered reasonable since noise is embedded in the collected samples. By applying the Bayesian theorem, the relevancies of different parameters to the fire accident was also analysed. The trained SVM equipped the general behaviour of the correlation between the input and output parameters of this study. It was applied with Monte Carlo approach to evaluate the relevancies. The results show that the Special-risk-building and Smoker are the two major parameters relevant to the fire accident. This result provides an important information to the NSW Government is plan their regulatory strategies for alleviate the effects causing fire accident.

Acknowledgements

The work described in this paper was fully supported by a grant from the Research Grants Council of the Hong Kong Special Administrative Region, China [Project No. CityU 116613].The authors would also like to acknowledge the support from the Fire and Rescue New South Wales (F&RNSW) to provide the useful fire records for this study and their professional comments on this statistical analysis.

References

[1] Vapnik, V.N., 1999. The Nature of Statistical Learning Theory. Springer-Verlag, New York, Inc.
[2] Suykens, J.A.K., 2001. Support Vector Machines: A nonlinear modelling and control perspective. European Journal of Control 7(2-3), p. 311-327.