Prediction of dwelling fire occurrences using data mining technologies

L. Yang¹, M. R. Brown¹, C. W. Dawson², M. Gell¹
¹School of Computing and Technology, University of Derby, UK
²Department of Computer Science, Loughborough University, UK

Abstract
Two data mining technologies have been employed in this paper to build a prediction model for dwelling fire incidents. The first method uses a feed-forward neural network to describe the relationship between the number of fires and factors that influence the fire occurrence. Eight factors are initially selected as the inputs of the neural network. Principle Component Analysis (PCA) is employed for pre-processing the input data set to reduce the number of the inputs. The model can give a prediction with an acceptable accuracy for the number of fires in dwelling areas. Genetic algorithms (GAs) are the second approach discussed in this article. The first three principle components of the raw data set chosen from outputs of the PCA are classified into the different groups according to their number of fires. An iterative GA is proposed and applied to extract the features for each data group. Once the features for all the groups have been identified the test data set can be easily clustered into one of the groups based on the group features. The number of fires for the group, which the test data belongs to, is the prediction of the fire occurrence for the test data. The two approaches have been compared in the conclusion.

1 Introduction
Fire, the oldest combustion problem, is a major cause of human suffering and material loss, and the one that perhaps we predict the least accurately (Fernandez-pello, 1994). A major reason for this is the complexity of the problem itself. Tremendous progress has been made in the last two
decades in developing fire models and the prediction of fire development, particularly in buildings (Fernandez-pello, 1994). However with regards to the fire occurrence prediction, most existing work focuses on prediction of the daily occurrence of human caused wildfires in forest or wild lands (Vega-Garcia et al., 1996; Martell et al., 1987; Chou et al., 1993). In order to minimize the threat of loss from dwelling fires, fire managers must be able to plan protection strategies that are appropriate for individual local areas. A prerequisite for this planning is the ability to predict for broad areas the local potential of fire incidences. Based on such geographic information, managers can establish priorities over the area for the allocation of suppression forces to improve the probability for initial attack to control fires that do occur in areas of high concern. Unfortunately, prediction of fire occurrence in dwelling areas has not been fully explored and remains as a great challenge to academics since the occurrence of dwelling fires is influenced by many human behaviour factors, such as smoking, cooking style, and alcohol consumption. In the authors’ knowledge the first work of fire prediction occurrence for dwelling areas in UK was done in 1971 by a research group in the Home Office (Home office, 1971). In their work, data were available on fire incidents in dwellings and other buildings, and estimates of residential and working populations in 1971 were obtained. Three statistical equations were built to describe the relationship between fire incidence and population, thus enabling predictions of future fire incidence to be calculated from projected populations. In the first equation, the fire incidence in a map grid square is assumed to be directly proportional to residential population and working population. In the other two equations population is assumed to be related to a transformed incidence variable. The transformations have the effect of causing incidence to increase more rapidly with increasing population. The incidence estimates were used in a study of fire station sitting in Northamptonshire in 1991.

In this paper, neural networks and GA technologies have been employed to build a prediction model for dwelling fire incidences. Statistics analysis has been used to select the input variables for the models. In order to optimise the number of the input variables in the neural network model Principal Component Analysis is employed to reduce the dimension of the input vector. An iterative GA is proposed and applied to extract the features for the training data groups. The number of fires for the training data group, which the test data belongs to in terms of the features extracted, is the prediction of the fire occurrence for the test data.

2 Determination of model inputs

The selection of appropriate model inputs is extremely important (Faraway and Chatfleld, 1998). The less superfluous information the model is given, the better it is able to latch on to the true relationships in the data. The choice of input variables is generally based on a prior knowledge of causal variables
in conjunction with inspections of time series plots of potential inputs and outputs.

Previous work (Home Office, 1971) determined that the residential population is the important factor in determining the number of dwelling fires and the working population is important in determining the number of non-dwelling fires. It has also been found that in each case the contribution of the second independent variable to the number of fires is statistically significant. The interpretation was: in areas with large residential and working populations the residential population is more densely housed and there is a disproportionate increase in fire incidence per head of population. However, the fire incidence in shops, schools and recreation centres would also be related to residential rather than working population.

Moving on from the previous work, the unemployment population has been considered as an input variable in this study rather than the working population. Furthermore, the population distributions in various ages are also selected as the inputs of the prediction model. The population in different ages can make different contributions to dwelling fire occurrence. For example, the houses where children and elder people are living may have a high probability of occurrence of fire incidence. People usually think the air temperature is one of the major factors to dwelling fire occurrence. The effect of the maximum and minimum air temperatures on the dwelling fire occurrence has been investigated. It was surprising that none of these drivers has a significant effect on the dwelling fire incidences. Later in this paper, correlation analysis and Chi-square test between these variables and the number of dwelling fires have proved the above assumptions.

It should be pointed out that human behaviour factors, such as smoking, chip pan cooking, drinking, and drug use, are thought of as having significant influence on dwelling fire occurrence. Unfortunately, data in these aspects are not available. Therefore they are not selected as the model inputs.

The correlation measures the direction and strength of the linear relationship between two quantitative variables. In order to analyse $n$ individual data on variable $x$ and variable $y$ the correlation $r$ between $x$ and $y$ is defined as

$$ r = \frac{1}{n-1} \sum \left( \frac{x_i - \bar{x}}{s_x} \right) \left( \frac{y_i - \bar{y}}{s_y} \right) $$

(1)

where $\bar{x}$ and $\bar{y}$ are the means of respectively $x_i (i=1,2,...,n)$ and $y_i (i=1,2,...,n)$, and $s_x$ and $s_y$ are the standard deviations. Table 1 shows the correlation analysis results between the number of dwelling fires and the population, unemployment, maximum temperatures and minimum temperatures. Obviously, there are strong correlations between the number of dwelling fires and the population and the unemployment levels, but there are very weak correlations between the number of dwelling fires and the maximum and minimum air temperatures. Therefore, the population and the unemployment levels have been selected as the inputs of the fire prediction model.

Similarly, Table 2 shows the analysis results for the population distributions among different ages. As discussed above the population of the children
group and elder people group have high correlations with the number of dwelling fires. Other groups also have a quite strong correlation with the fire number, and should not be ignored in the prediction model.

Table 1. Correlation between the number of dwelling fires and the statistic variables.

<table>
<thead>
<tr>
<th>Correlation</th>
<th>Population</th>
<th>Unemployment</th>
<th>Maximum temperature</th>
<th>Minimum temperature</th>
</tr>
</thead>
<tbody>
<tr>
<td>Number of dwelling fires</td>
<td>0.528</td>
<td>0.554</td>
<td>-0.044</td>
<td>-0.0306</td>
</tr>
</tbody>
</table>

Table 2. Correlation between the number of dwelling fires and the population distribution.

<table>
<thead>
<tr>
<th>Correlation</th>
<th>Population in various ages</th>
</tr>
</thead>
<tbody>
<tr>
<td>Number of dwelling fires</td>
<td>0~4</td>
</tr>
<tr>
<td></td>
<td>0.564</td>
</tr>
</tbody>
</table>

The Chi-square test is another statistic method that has been used for selecting the inputs. A similar result has been obtained. The detail is omitted here.

3 Neural network based prediction model

The objective of building the prediction model for dwelling fires is to help fire managers to position their limited resources appropriability. The dwelling fires are best controlled when they are small. Sound fire management therefore demands that initial attack resources be deployed so they can be quickly dispatched to fires as they are reported. Fire managers must conduct thorough deployment analyses to ensure their limited resources are well positioned before they are actually needed. In order to accomplish that task they require accurate fire occurrence predictions for particular areas.

The Derbyshire area was divided into 8 districts and further into 189 wards. Data were obtained from various sources on population, unemployment levels, the population distribution in various ages, the air temperatures and the number of dwelling fires in each ward in the year 1998.

The number of dwelling fires is selected as the output variable of the neural network for the ward based prediction model, the input variables of the neural network are the unemployment and the population distribution in the various ages rather than the population itself. The neural network takes the following forms:
where FIRE denotes the number of dwelling fires in a ward, AGE0 the population under 4 years old in the ward, AGE5 the population from 5 to 11, AGE12 the population from 12 to 18, AGE19 the population from 19 to 64, AGE65 the population from 65 to 74, AGE75 the population over 75 years old, UNEMP the unemployment. NN represents a neural network mapping. In order to optimise the number of the inputs the PCA is employed for dimensionality reduction of the input vectors and generation of orthogonal variables. In the PCA, the covariance matrix is decomposed in terms of its $p$ eigenvectors and then orders the resulting principal components so that those with the largest variation come first. Only the first $r$ eigenvectors (associated with the larger eigenvalues) are retained, while $(p-r)$ smaller components are discarded, assuming that the latter describe mostly noise. The selection of the eigenvalues to be discarded is often a difficult task and various methods may yield completely different results, as was recently demonstrated by Valle et al (1999). In this study the trial and error method has been used for the selection of the eigenvalues to be discarded. Figure 1 shows the variances that all the principal components (PCs) captured for all the input variables in the model NN. From Figure 4 it can be seen that PC1, PC2, and PC3 capture the 95.5% variance for all the input variables, but PC4 to PC7 only capture 4.5% variance. If the PCA is set to eliminate those principal components which contribute less than 5% to the total variation in the data set, the number of inputs will be reduced from 7 to 3 after using PCA as an input filter. Exploratory analysis was used to select the major suitable transfer functions (sigmoid, log-sigmoid, and/or linear transfer functions) and learning algorithms (Backpropagation algorithm, Levenberg-Marquardt algorithm, or Bayesian automated regularization). Trials indicated that the best results for this problem could be achieved by using the sigmoid transfer function at the hidden layers and the linear transfer function at the output layer, and the Levenberg-Marquardt algorithm. Different network architectures were tested, yet all trials employed one and two hidden layers. The best result was obtained from a network with two hidden layers and 10 nodes in the first hidden layer and 25 nodes in the second hidden layer. The data set was divided into two parts, 8 randomly selected wards, the wards 4, 7, 15, 76, 90, 98, 110, and 122, acting as the test set, and the rest data acting as the training set. The training results don’t have any error between the actual output and the predicted output. The test results are shown in Figure 2. The maximum error is 3 fires (the actual fire number is 6, but the predicted fire number is 9) for the ward 15. The predicted errors in another wards are 0 (the ward 90), 1 (the wards 4, 7, 76, 98, 110), and 2 (the wards 122). The mean squared error (MSE) is selected to assess the results as follows. The MSE is 2.375 for this set of test data.
Figure 1: Variance explained by the principal components.

Figure 2: Test results of the neural network model.

\[ MSE = \frac{\sum_{i=1}^{n} (A_i - M_i)^2}{n} \]  

where \( A_i \) are the actual number of dwelling fires and \( M_i \) are the predicted number of dwelling fires, \( n \) is the number of the test data set.
4 GA based prediction model

4.1 Methodology

The statistics of dwelling fire incidence in the Derbyshire area in the past decade shows that the maximum number of dwelling fires in individual wards is normally less than 15. In 1998 the maximum number is 12. Therefore the original data set of dwelling fire incidents can be partitioned into several groups in terms of the number of fires. A number of data mining technologies are available for extracting features for the individual groups such as GAs. Once the features for all the groups have been identified the test data set can be easily cast into one of the groups based on the group features. The number of fires for the group, which the test data belongs to, is the prediction of the fire occurrence for the test data.

Similar to the neural network approach described above, in order to reduce the data set dimension PCA has been used as a data filter for the raw data set. Only the first three principal components (PC1, PC2, and PC3 shown in Figure 1) are used for group feature extraction because they have captured 95.5% variance for all the input variables. The raw data set is transferred to a data set in the new three-dimensional coordinate system defined by the PC1, PC2 and PC3.

4.2 Features extraction by using an iterative GA

The features for each group can be represented by a set of guide points (Adriaans and Zantinge, 1996). Therefore feature extraction has been transferred into a clustering problem for searching the guide points. GAs provide a good mechanism to find the guide points. Suppose that k guide points $P_1, P_2, \ldots P_k$ are employed and the string comprising the three-dimensional coordinates of all guide points is chosen as a chromosome. The coordinate of the guide point $P_i$ ($i=1, 2, \ldots, k$) is denoted as $(X_{pi}, Y_{pi}, Z_{pi})$. The chromosome is expressed as a real string with a length of $3k$ (the triple of the number of the guide points) as below:

$$X_{p1} \ Y_{p1} \ Z_{p1} \ \ldots \ X_{pk} \ Y_{pk} \ Z_{pk} \ (4)$$

A generation consists of a number of chromosomes (50 in this study). In the initial stage, all guide points will be selected randomly, so the GA will start with several chromosomes that describe a number of random solutions to the clustering problem. The average distance of the points in the group to the closest guide point is chosen as a fitness function as below. A set of guide points will give a better clustering if the average distance of all the points to the closest guide point is minimal. Taking the number of empty clusters into consideration, the fitness function includes a term to penalize degenerate solution (Hall et al., 1999; Meng et al., 2000). Any value of the fitness function is scaled with a penalty factor. This penalty factor is represented by
1 plus the number of the empty clusters, E, divided by the total number of clusters $k$.

$$f_{\text{fitness}} = \frac{1}{N} \sum_{i=1}^{N} \min_{j=1}^{k} \left( (X_i - X_{pj})^2 + (Y_i - Y_{pj})^2 + (Z_i - Z_{pj})^2 \right)^{1/2} (1 + E / k)$$

Since there is not any pre-knowledge about the number of guide points available, an iterative GA is proposed here.

As shown in Figure 3, the number of guide points, $k$, is initialised as 1. The single GA execution is terminated by the maximum number of generations or the fitness value if the desired value is achieved. If the desired fitness value is not achieved by the termination of the GA the number of guide points $k$ increments by 1, the next iteration starts and the GA is invoked again. The procedure repeats until the desired fitness value is achieved. The desired fitness value is chosen by using the trial and error method.

4.3 Results

From experiments, 14 guide points have been obtained by using the iterative GA for 8 groups (the fire numbers of 10 and 12 have not been counted in as there is only one ward in correspondence with each of them). The groups with the number of fires from 0 to 4 have two guide points each. The rest groups with the number of fires from 5 to 7 have a single guide point each. Prediction can be carried out based on the set of the guide points obtained. As described in the methodology, if any test data belongs to one of the groups the number of fires for the group is the prediction of the fire occurrence for the test data. The 8 randomly selected wards for the neural network approach, the wards 4, 7, 15, 76, 90, 98, 110, and 122, are still used as the test set in the
GA approach. The prediction results for these test data is shown in Figure 4. The MSE is 2.875 for this set of test data, which is very close to the result of the neural network approach.

![Graph showing prediction results](image)

Figure 4: Test results of the GA based prediction model.

5 Conclusions

It is well known that data are the most important factor in determining a model’s performance. Any model is only as good as the data used in developing it. Data available for model building in this study are very limited in quality and quantity. Some variables, which may have significant influences on the occurrence of the dwelling fires, are omitted in the prediction model, not because they are less related, but because the data about these variables are not available. These variables include the smoking population, the chip pan cooking population, the education level, the drug using population, and the population of over use of alcohol. The internal report of the Derbyshire Fire & Rescue Service has shown that smoking, chip pan, misuse of electrical equipment and too close to heat sources are the main causes of fire in domestic properties. The lack of these important data is one of the main reasons why the prediction results are not as good as expected.

In comparison to more traditional models such as linear and/or non-linear regression models, neural network models have both advantages and disadvantages. If the training data over the whole possible range are available the neural network model may provide a prediction with a high accuracy. But
the computational costs were high in producing the networks. More important, fire managers are less likely to accept a “black box” systems such as the ones built above when a more familiar and better understood procedure such as a regression analysis is available.

The GA approach proposed in this article provides an alternative way for the dwelling fire prediction. Comparing with the neural network model this method is more robust and not sensitive to the test data may because of the novel nature of the method.

Two interesting conclusions can be made at the end of this paper. Air temperature has little correlation with the number of dwelling fires, which is completely opposite to common opinion. The number of dwelling fires can be properly modelled and accurately predicted by using neural networks and GA technologies only if the relevant data are available.

Acknowledgement

This work was financially funded by Derbyshire Fire & Rescue Service.

References