Expert Systems with Applications 37 (2010) 7729–7736

Contents lists available at ScienceDirect

Expert Systems with Applications

journal homepage: www.elsevier.com/locate/eswa

# A decision making system to automatic recognize of traffic accidents on the basis of a GIS platform

S. Savaş Durduran *

*Selcuk University, Engineering of Faculty, Department of Geomatic Engineering Campüs/Konya, Turkey*

ARTICLE INFO

*Keywords:*
Geographical information systems (GIS)
Traffic accident analysis
Decision making system
Correlation-based feature selection
Support vector machine
Artificial neural network

ABSTRACT

The prediction of traffic accidents is one of most important issues in our life. In the prediction of traffic accidents, a GIS platform to extract the important features including day, temperature, humidity, weather conditions, and month of occurred traffic accidents has been used. In this study, a decision making system (DMS) based on correlation-based feature selection and classifier algorithms including support vector machine (SVM) and artificial neural network (ANN) has been proposed to predict the traffic accidents identifying risk factors connected to the environmental (climatological) conditions, which are associated with motor vehicles accidents on the Konya–Afyonkarahisar highway with the aid of geographical information systems (GIS). Locations of the motor vehicle accidents are determined by the dynamic segmentation process in ArcGIS 9.0 from the traffic accident reports recorded by District Traffic Agency. In this DMS, firstly the number of dimension of traffic accidents dataset with five features (ay, temperature, humidity, weather conditions, and month of occurred traffic accidents) has been reduced from 5 to 1 feature by using correlation-based feature selection (CFS). In CFS method, the correlation coefficients between five features and outputs (the cases of without accident or with accident) has been calculated and chosen the feature that has highest correlation coefficient. Secondly, the traffic accident cases with one feature have been classified as without accident or with accident using SVM and ANN models. The proposed DMS has obtained the prediction accuracy of 61.79% with ANN classifier and achieved the prediction accuracy of 67.42% using SVM with RBF (radial basis function) kernel. These results have indicated that the proposed DMS could be used on prediction of real traffic accidents.



## 1. Introduction

The quantity of traffic accidents whose moral and material losses have reached to unbearable levels, increase day by day and the locations of these accidents are very important (Müge & Şenkal, 1999). Today, the most negative result of developing transportation systems is traffic accidents with injuries and loss of lives. The tremendous social and economic costs associated with traffic accidents have led many road authorities and researchers to establish safety management programs that aim to continually improve the safety performance of highways (Sawalha & Sayed, 2006). So, traffic safety is the most critical issue in agencies' transportation strategy. The identification of safety deficient areas on the highways is aimed to implement precautionary measures and provisions by researchers and traffic officials (Erdogan, Yilmaz, Baybura, & Gullu, 2007).

Geographical information system (GIS) technology is becoming an increasingly popular tool for visualization and analyses of accident data in highways. GIS has the ability to hold a vast amount of data that can be easily stored, shared analyzed and managed. It provides a platform for spatial data analyses and visualization to explore relationships between spatial and non-spatial data (Erdogan et al., 2007).

The success of safety improvement programs in reducing accident occurrence depends on the methods used in the accident analyses. Today, many researchers were used different deterministic and statistical methods in the studies that aiming to determine the high rate accident locations and safety deficient areas on the highways (Erdogan et al., 2007; Levine, Kim, & Lawrence, 1995; Loo, 2006).

The purpose of the road analysis is to determine the dangerous road sections, to take precautions that are suitable for those sections in order to prevent the accidents and diminish losses. The first stage in diminishing losses due to accidents is to determine the section where precaution needs to be taken. The suitable precaution can be taken only after the correct determination of this section. However, in order to make this determination correctly, a determination method that is suitable to the traffic conditions should be used. Every country uses a method that is suitable to

* Tel.: +90 332 223 1909; fax: +90 332 241 0635.
*E-mail address:* durduran@selcuk.edu.tr

0957-4174/$ - see front matter © 2010 Elsevier Ltd. All rights reserved.
doi:10.1016/j.eswa.2010.04.068

itself in order to determine the dangerous sections (Müge & Şenkal, 1999). Many of these researches have explored the relationships between traffic accidents and geometric design and operation of road segments. However, data relating to accidents are widely available, but have received surprisingly little analyses with respect to weather. The complexity involved in establishing the exact cause-and-effect relationship in traffic accidents acts as an obstacle, because road accidents are the results of an intricate driver–vehicle–environment matrix (Andreescu & Frost, 1998). So, it is aimed to decrease the accidents determining the effects of weather and some environmental conditions on traffic accidents in Konya–Afyonkarahisar highway with the aid of GIS and artificial intelligence.

A novel decision making system based on correlation-based feature selection and classification algorithms including support vector machine and artificial neural network for predicting the traffic accidents Konya–Afyonkarahisar highway in Turkey with the aid of GIS. The used traffic accidents dataset comprises five features including day, temperature, humidity, weather conditions, and month of occurred traffic accidents and comprises 378 data points (179 without traffic accident and 179 with traffic accident). In order to select the significant feature from dataset and to reduce the complexity of classification algorithms, the correlation-based feature selection method has been applied to traffic accidents dataset and selected the first feature of dataset that is most related to class cases of traffic accidents dataset in the end of this process. After feature selection process, the traffic accidents dataset with one feature (day attribute) has been assigned to either without accidents or with accident via SVM or ANN classifiers. Thanks to this decision making system, the case of traffic accident according to obtained information from a GIS could be determined.

The rest of the paper is arranged as follows. The material is described in the next section. Section 3 presents the proposed decision making system. The experimental data and results to present the effectiveness of proposed method are given in Section 4. The conducted conclusions and discussions are given in Section 5.

## 2. Materials

### 2.1. Traffic accident dataset

Afyonkarahisar–Konya highway is a junction region in Turkey connecting the industrial, tourism and agricultural areas to each others. Especially in winter, weather related crashes happen frequently because of continental climate in the region (see Fig. 1). The length of the highway is approximately 240 km 65% of the road has two lanes. Traffic accident reports belong the highway are obtained on paper form by District Traffic Agency Officers in Turkey. These records include collected accident parameters such as the date, hour/minutes, kilometer of crash, code of highway, age, sex and alcohol consumption of driver, weather conditions, lighting conditions, vehicle type, and number of persons injured/killed. Table 1 presents the statistical values of attributes of raw traffic accident dataset. The important thing here is to select the correct database and keeping the accident data correct, updated and complete. In the database, the effects of road situation, environmental conditions and the vehicle status, on the accident will be determined. Besides, coordinates will be determined by GPS and the spatial situation will be correlated with the map.

Using this accident data a GIS-based study performed for the purpose of reducing the number of the accidents by determining the effects of environment and weather conditions on the phenomenon of accidents. First, 179 accident and 179 not accident records belong to 2006 year and were collected from the District Traffic Agency and input into an MS Access database. Meanwhile, highway was digitized at a scale of 1:1,000,000 with ArcGIS 9.0 software. The location of the accidents positioned with the "kilometer of crash" data on the route of highways using the "linear referencing" toolset in ArcGIS 9.0. Making the accident analysis in the region by taking advantage of the database created by the correlation of regional numerical map and accident data, provides a good infrastructure for the prevention of the accidents.

**Table 1**
The statistical values of attributes of raw traffic accident dataset.

| Attribute number | Minimum | Maximum | Mean | Standard deviation |
|---|---|---|---|---|
| 1 | 1 | 7 | 4.23 | 1.97 |
| 2 | −4 | 29 | 14.36 | 8.31 |
| 3 | 20 | 91 | 53.53 | 21.74 |
| 4 | 1 | 5 | 2502 | 1.29 |
| 5 | 1 | 12 | 7.74 | 3.16 |


**Fig. 1.** Konya–Afyonkarahisar highway in Turkey.

## 3. Methods

### 3.1. The proposed decision making system

In this paper, a decision making system comprising correlation-based feature selection and support vector machine or artificial neural network classifier algorithms has been proposed to predict the traffic accidents on Afyonkarahisar–Konya highway in Turkey. The proposed method comprises two stages. The first stage is feature selection process. In this stage, the correlation-based feature selection to select the significant features from dataset was used. The second stage is the classification process. In this stage, the SVM and ANN classifier algorithms have been used to classify the traffic accidents with selected features. Fig. 2 presents the block diagram of proposed DMS.

### 3.2. Correlation-based feature selection (CFS): feature selection process

Correlation is a criterion used measuring whether a feature (attribute) is the relevant to other features in dataset and the relevant to outputs (classes) or not. With respect to this idea, we have proposed a feature selection method based on correlation coefficients between features incorporating dataset and the outputs of dataset. A feature is good if it is relevant to the class however is not redundant to any of the other relevant features. Applied with correlation, the goodness of feature is measured whether it is highly correlated with the class but not highly correlated with any of the other features (http://www.mathbits.com/Mathbits/TISection/Statistics2/correlation.htm; www.mathworks.com).

The linear correlation coefficient ($R$), measures the strength and the direction of a linear relationship between two variables is given as follows (www.mathworks.com):

$$R(i,j) = \frac{C(i,i)}{\sqrt{C(i,i)C(j,j)}} \quad (1)$$

where $R$ value is the correlation coefficient between variables $i$ and $j$, $C(i,j)$ is the covariance matrix and is given in Eq. (2):

$$\mathrm{COV} = \frac{\sum_{i=1}^{n}(X_i - \bar{x})(Y_i - \bar{y})}{n-1} \quad (2)$$

In above, the formula of covariance (COV) of the variables $X$ and $Y$ is given. Where $\bar{x}$ and $\bar{y}$ denote the means of $X$ and $Y$, respectively. And also, $n$ is the number of pairs of data.

The value of $R$ varies between $-1$ and +1. The + and $-$ signs are used for positive linear correlations and negative linear correlations, respectively. The three types of correlations between variables are explained as follows (http://www.mathbits.com/Mathbits/TISection/Statistics2/correlation.htm; www.mathworks.com):

*Positive correlation: if* $x$ and $y$ have a strong positive linear correlation, $R$ is close to +1. An $R$ value of exactly +1 indicates a perfect positive fit.

*Negative correlation: if* $x$ and $y$ have a strong negative linear correlation, $R$ is close to $-1$. An $r$ value of exactly $-1$ indicates a perfect negative fit.

*No correlation: if* there is no linear correlation or a weak linear correlation, $R$ is close to 0.

The correlation is a dimensionless quantity; that is, it does not depend on the units employed. A perfect correlation of ±1 happens only when the data points lie exactly on a straight line. If $R = +1$, the slope of this line is positive. If $R = -1$, the slope of this line is negative (http://www.mathbits.com/Mathbits/TISection/Statistics2/correlation.htm; www.mathworks.com).

In calculating of correlation coefficients, the $p$-value is computed by transforming the correlation to create a $t$ statistic having $N-2$ degrees of freedom, where $N$ is the number of rows of $X$. If $p(i,j)$ is small, say less than 0.05, then the correlation $R(i,j)$ is significant (www.mathworks.com).

Fig. 3 demonstrates the flowchart of correlation-based feature selection. And also, Table 2 shows the correlation coefficients between features forming traffic accidents dataset with class case.

As can be seen from Table 2, the first feature (day feature) is a significant feature since its $p$-value is smaller than value of 0, 0.5. So, we have only chosen the first feature of traffic accidents dataset in experimental studies.

### 3.3. Used classifier algorithms: classification process

After CFS process, the classifier algorithms including artificial neural network and SVM classifiers are used. These algorithms have been explained in the following subsections. In the training and testing of classifiers, 50–50% training–testing dataset split is used.

#### 3.3.1. Support vector machine (SVM)

SVM is a reliable classification technique, which is based on the statistical learning theory. This technique was firstly proposed for classification and regression tasks by Vapnik (1995).

It is a method for creating functions from a set of labeled training data. The function can be a classification function (the output is binary: is the input in a category) or the function can be a general regression function. For classification, SVMs operate by finding a hypersurface in the space of possible inputs. This hypersurface will attempt to split the positive examples from the negative examples. The split will be chosen to have the largest distance from the hypersurface to the nearest of the positive and negative examples. Intuitively, this makes the classification correct for testing data that is near, but not identical to the training data (http://research.microsoft.com).

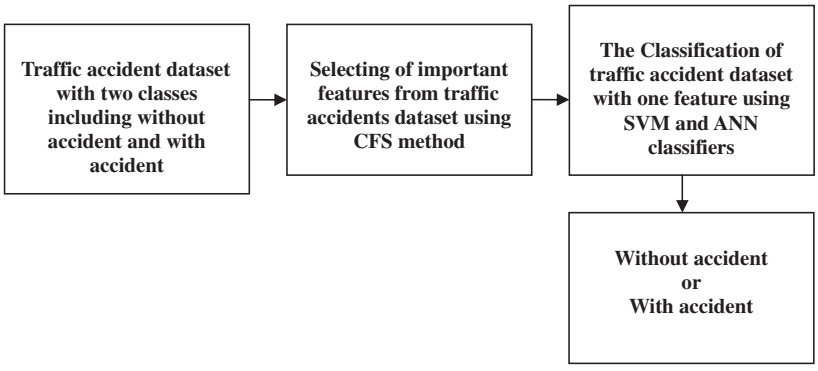


**Fig. 2.** The block diagram of proposed DMS.

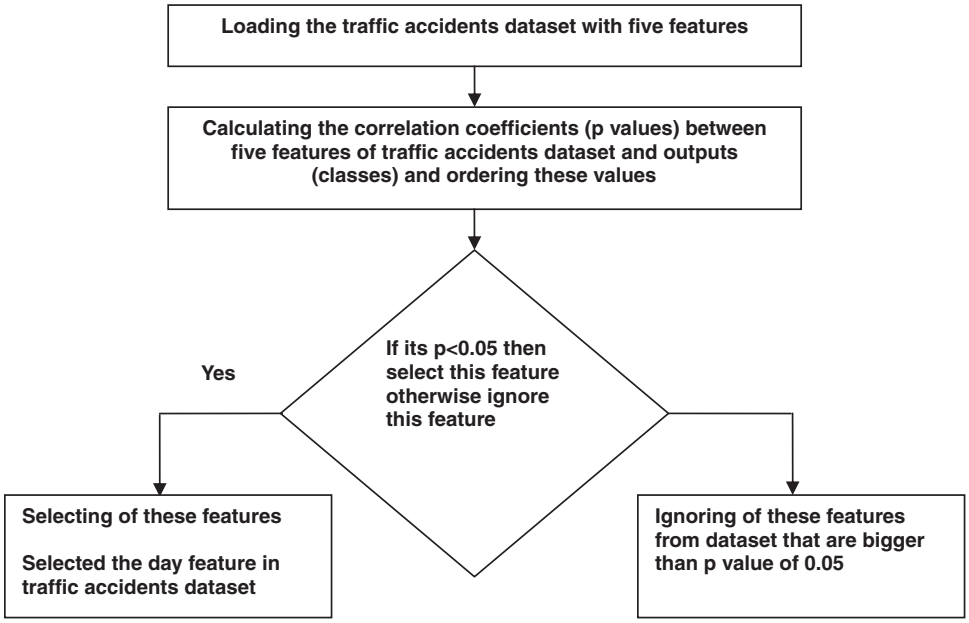


**Fig. 3.** The flowchart of correlation-based feature selection.

**Table 2**
The correlation coefficients between features forming traffic accidents dataset with class case.

| Number of features in dataset | Correlation coefficient between feature and output (class case) | *p*-Value |
|---|---|---|
| 1 (day) | 0.16846 | 0.0013 (<0.05) |
| 2 (temperature) | 0.04340 | 0.4128 |
| 3 (humidity) | 0.02456 | 0.6431 |
| 4 (weather conditions) | −0.0517 | 0.3291 |
| 5 (month of occurred traffic accidents) | 0.10346 | 0.0504 |

In literature, there are a lot of works related to applications of SVM in the pattern recognition and biomedical engineering fields. Among these, Comak et al. diagnosed the heart valve diseases using support vector machine as a biomedical application (Comak, Arslan, & Türkoğlu, 2007). Polat et al. used the least square support vector machine (LSSVM) classifier to diagnose the breast cancer disease (Polat & Güneş, 2007). Huang et al. used three strategies to construct the hybrid SVM-based credit scoring models to evaluate the applicant's credit score from the applicant's input features (Huang, Chen, & Wang, 2007). Camastra presents a cursive character recognizer, a crucial module in any cursive word recognition system based on a segmentation and recognition approach using SVM (Camastra, 2007).

As shown in Fig. 4, a linear SVM was developed to classify the data set which contains two separable classes such as {+1,−1}. Let the training data consist of n datum $(x_1,y_1),\ldots,(x_n,y_n)$, $x \in R^n$ and $y \in \{+1,-1\}$. To separate these classes, SVMs have to find the optimal (with maximum margin) separating hyperplane so that SVM has good generalization ability. All of the separating hyperplanes are formed with

$$D(x) = (w \times x) + w_0 \tag{3}$$

and provide following inequality for both $y = +1$ and $-1$

$$y_i[(w \times x_i) + w_0] \geqslant 1, \quad i = 1,\ldots,n \tag{4}$$

The data points which provide above formula in case of equality are called the support vectors. The classification task in SVMs is implemented by using these support vectors.

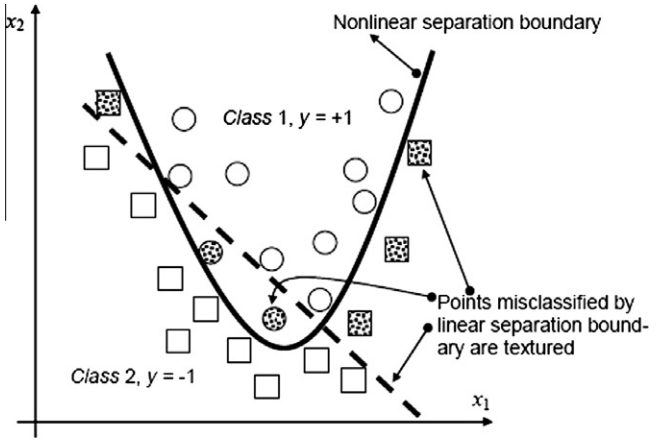


**Fig. 4.** The basic structure of a simple SVM.

Margins of hyperplanes comply with following inequality

$$\frac{yk \times D(xk)}{||w||} \geqslant \Gamma, \quad k = 1,\ldots,n \tag{5}$$

To maximize this margin ($\Gamma$), norm of $w$ is minimized. To reduce the number of solutions for norm of $w$, following equation is determined

$$\Gamma \times ||w|| = 1 \tag{6}$$

Then formula (7) is minimized subject to constraint (4)

$$1/2||w||^2 \tag{7}$$

When we study on the non-separable data, slack variables $\xi_i$, are added into formula (4) and (7). Instead of formulas (4) and (7), new formulas (8) and (9) are used

$$y_i[(w \cdot x_i) + w_0] \geqslant 1 - \xi_i, \tag{8}$$

$$C\sum_{i=1}^{n} \xi_i + 1/2||w||^2 \tag{9}$$

Since originally SVMs classify the data in linear case, in the nonlinear case SVMs do not achieve the classification tasks. To over-

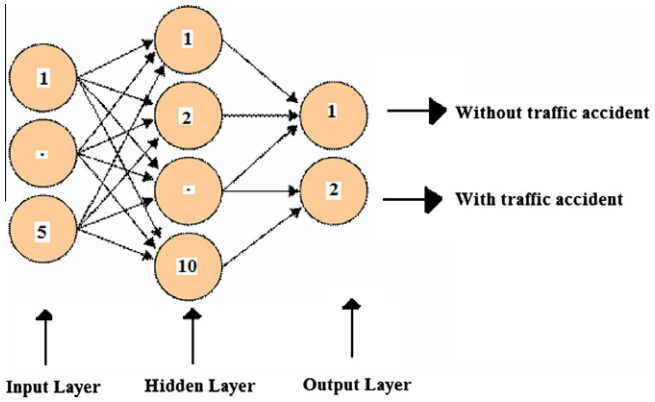


**Fig. 5.** The structure of ANN with LM.

**Table 3**
The number of nodes in the input, hidden, and output layers on network.

| Dataset | The number of node in input layers | The number of node in hidden layers | The number of node in output layers |
|---|---|---|---|
| Traffic accident dataset | 5 | 10 | 2 |

come this limitation on SVMs, kernel approaches are developed. Non-linear input data set is converted into high dimensional linear feature space via kernels. In SVMs, following kernels are most commonly used.

- Dot product kernels: $K(x,x') = x \cdot x'$.
- Polynomial kernels: $K(x,x') = (x \cdot x' + 1)^d$; where $d$ is the degree of kernel and positive integer number.
- RBF kernels: $K(x,x') = \exp(-||x - x'||^2/\sigma^2)$; where $\sigma$ is a positive real number.

In SVM classifier, these three kernel functions have been used and compared to each other. And then the best kernel function has been selected. In the RBF kernel function, valued of $\sigma$ and $C$ are selected as 10 and 100 using trial and error method, respectively.

*3.3.2. Artificial neural network (ANN) with levenberg Marquart*

An ANN is constructed for a specific application, such as pattern recognition or data mining, by means of a learning process. The back propagation (BP) algorithm is a most widely used training procedure that adjusts the connection weights of a multi layer perceptron (MLP) (Haykin, 1999). Mainly, the LM algorithm is a least-squares estimation algorithm on the basis of the maximum neighborhood idea. A MLP comprises of three layers: an input layer, an output layer, and one or more hidden layers. Each layer comprised of a determined number of neurons. The neurons in the input layer only work as buffers for distributing the input signals $\mathbf{x}_i$ to neurons in the hidden layer (Haykin, 1999; Lawrence, 1994).

In applications, the number of node in input layers and the number of node in output layers change according to the number of samples and class labels in used dataset, respectively. Fig. 5 shows the structure of ANN with LM. Table 3 presents the number of node in the input, hidden, and output layers on network in prediction of traffic accident dataset using MLPANN.

## 4. Results and discussion

In this section, the performance measure criteria to test the performance of proposed method have been given. Later, the subsection of results and discussion has been explained.

*4.1. Performance measurement criteria*

In order to test the performance of proposed method, the classification accuracy and sensitivity–specificity values were used on the predicting of traffic accidents in Konya–Afyonkarahisar highway in Turkey. These measures were shortly explained on a confusion matrix. After constructing our model and testing it using 50–50% train–test split, we build the confusion matrix for the dataset. Confusion matrix is shown in Table 4 (actual vs. predicted) and the other parameters which are computed using confusion matrix are shown with the following equations (Lewis, 1994; Ma, Guo, & Cukic, 2006).

The entries in the confusion matrix have the following meaning in the context of our study:

- TN is the number of correct predictions that an instance is negative.
- FP is the number of incorrect predictions that an instance is positive.
- FN is the number of incorrect of predictions that an instance negative.

**Table 4**
Representation of confusion matrix.

| Actual | Predicted | |
|---|---|---|
| | Negative | Positive |
| Negative | TN | FN |
| Positive | FP | TP |

**Table 5**
The obtained results from ANN and SVM classifiers without correlation-based feature selection using 50–50% training–test partition on the classification of traffic accidents with five features in a GIS platform.

| Classifier algorithm | Kernel type | Classification accuracy (%) | Sensitivity (%) | Specificity (%) |
|---|---|---|---|---|
| Artificial neural network | – | **53.93**[a] | **52.67**[a] | **57.44**[a] |
| Support vector machine | RBF kernel function | 52.25 | 51.39 | 55.88 |
| | Polynomial kernel function | 37.64 | 37.78 | 37.50 |
| | Linear kernel function | 48.31 | 0.0 | 49.14 |

[a] Indicates the obtained best classification results on prediction of traffic accidents.

**Table 6**
The obtained results from ANN and SVM classifiers using 50–50% training–test partition on the classification of traffic accidents with one feature (day feature of traffic accident dataset) selected by CFS in a GIS platform.

| Classifier algorithm | Kernel type | Classification accuracy (%) | Sensitivity (%) | Specificity (%) |
|---|---|---|---|---|
| Artificial neural network | – | 61.79 | 58.97 | 67.21 |
| Support vector machine | RBF kernel function | **67.42**[a] | **65.66**[a] | **69.62**[a] |
| | Polynomial kernel function | 62.92 | 59.82 | 68.85 |
| | Linear kernel function | 50.00 | 0.0 | 50.00 |

[a] Indicates the obtained best classification results on prediction of traffic accidents.

- TP is the number of correct predictions that an instance is positive.

Sensitivity and specificity are the most widely used statistics used to describe a diagnostic test. Sensitivity is the proportion of people that tested positive of all the positive people tested; that is (true positives)/(true positives + false negatives). It can be seen as the probability that the test is positive given that the patient is sick. The higher the sensitivity, the fewer real cases of diseases go undetected (or, in the case of the factory quality control, the fewer faulty products go to the market). Specificity is the proportion of people that tested negative of all the negative people tested; that is (true negatives)/(true negatives + false positives). As with sensitivity, it can be looked at as the probability that the test is negative given that the patient is not sick. The higher the specificity, the fewer healthy people are labeled as sick (or, in the factory case, the less money the factory loses by discarding good products instead of selling them) (http://en.wikipedia.org/wiki/Binary_classification).

For sensitivity and specificity analysis, we use the following expressions.

*Sensitivity*: it is computed using Eq. (10)

$$\text{Sensitivity} = \text{TP}/(\text{TP} + \text{FN}) \tag{10}$$

*Specificity*: it is computed using Eq. (11)

$$\text{Specificity} = (\text{TN})/(\text{TN} + \text{FP}) \tag{11}$$

*Classification accuracy*: accuracy is the likelihood of correctly predicted total number of modules and it is computed using Eq. (12)

$$\text{Accuracy} = (\text{TP} + \text{TN})/(\text{TN} + \text{FP} + \text{FN} + \text{TP}) \tag{12}$$

**Table 7**
The obtained confusion matrixes for ANN, SVM, combining of correlation-based feature selection and ANN, and combining of correlation-based feature selection and SVM on the recognition of traffic accidents in a GIS platform.

| Output/desired | Result (without accident) | Result (with accident) | Classifier methods |
|---|---|---|---|
| Result (without accident) | 27 | 62 | ANN classifier |
| Result (with accident) | 20 | 69 | |
| Result (without accident) | 19 | 70 | SVM classifier (RBF kernel) |
| Result (with accident) | 15 | 74 | |
| Result (without accident) | 41 | 48 | Combination of correlation-based feature selection and ANN |
| Result (with accident) | 20 | 69 | |
| Result (without accident) | 55 | 34 | Combination of correlation-based feature selection and SVM classifier (RBF kernel) |
| Result (with accident) | 24 | 65 | |

### 4.2. Results and discussion

The detection of traffic accidents is one of most significant subjects in our life. In the prediction of traffic accidents, a GIS platform to obtain the significant features containing day, temperature, humidity, weather conditions, and month of occurred traffic accidents was used. In this study, a decision making system on the basis of correlation-based feature selection and classifier algorithms including support vector machine and artificial neural network has been suggested to classify the traffic accidents identifying risk factors connected to the environmental conditions, which are associated with motor vehicles accidents on the Konya–Afyonkarahisar highway with the aid of geographical information systems (GIS). The locations of the motor vehicle accidents are determined via the dynamic segmentation process in ArcGIS 9.0 from the traffic accident reports recorded by District Traffic Agency. In SVM classifier, three different kernel functions including RBF kernel, polynomial, and linear kernel functions have been used and compared to each other. Table 5 gives the obtained results from SVM and ANN classifiers without correlation-based feature selection on the prediction of traffic accidents using 50–50% training–testing split of whole dataset.

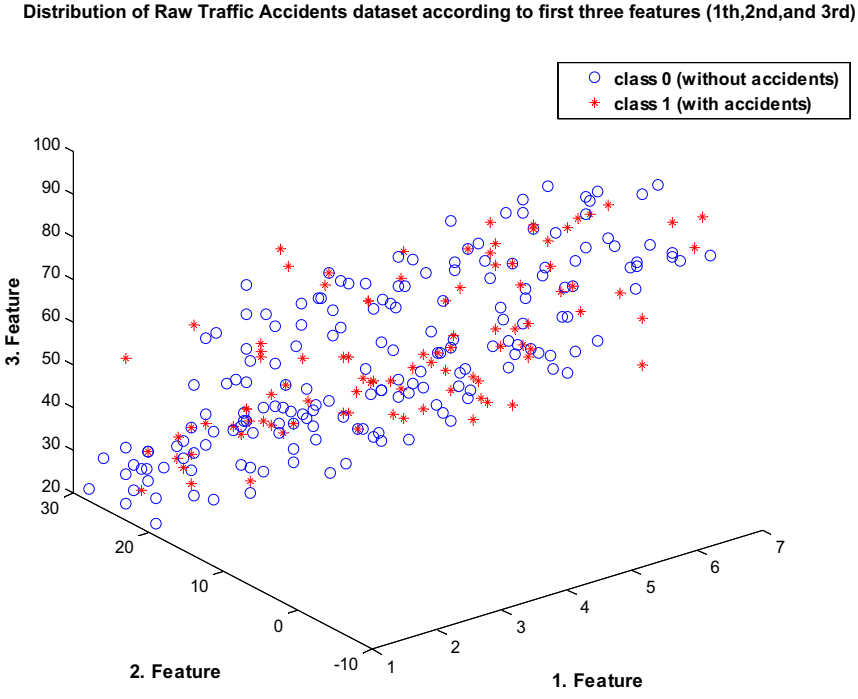


**Fig. 6.** The distribution of raw traffic accident dataset according to first three features (1st, 2nd, and 3rd attributes).

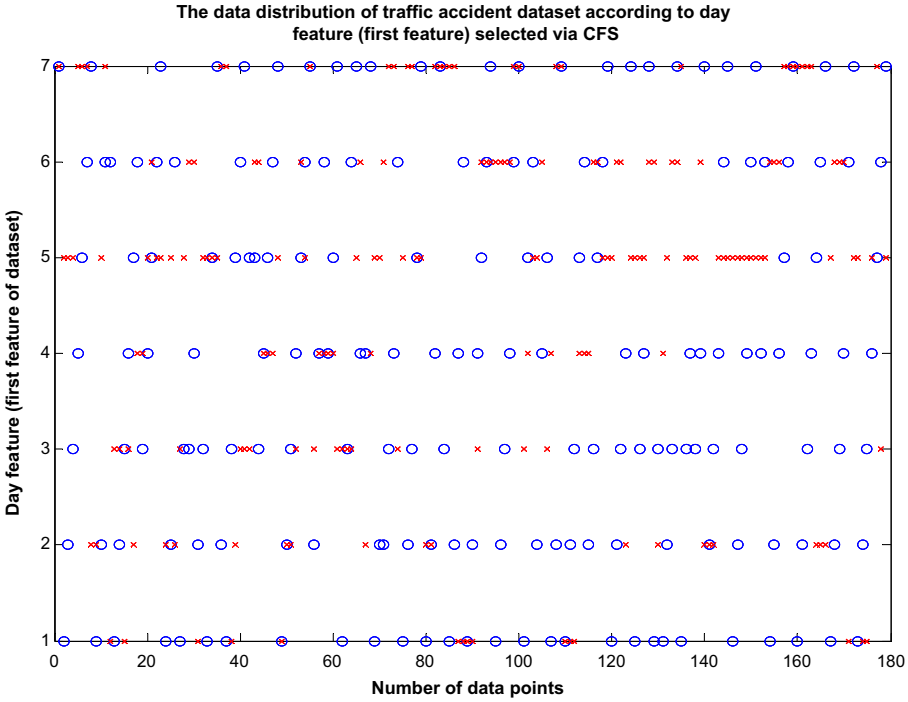


**Fig. 7.** The data distribution of raw traffic accident dataset with one feature (day feature) selected by CFS.

Table 6 provides the obtained results from SVM and ANN classifiers with correlation-based feature selection on the prediction of traffic accidents using 50–50% training–testing split of whole dataset.

Table 7 displays the confusion matrixes for ANN, SVM, combining of correlation-based feature selection and ANN, and combining of correlation-based feature selection and SVM on the recognition of traffic accidents in a GIS platform.

In order to further see the effect of correlation-based feature selection (CFS), the data distributions of traffic accidents dataset with first three features and one feature selected by CFS have been provided. Fig. 6 presents the distribution of raw traffic accident dataset according to first three features (1st, 2nd, and 3rd attributes). Fig. 7 demonstrates the data distribution of raw traffic accident dataset with one feature (day feature) selected by CFS.

Thanks to feature selection process, the non-linearly separable traffic accidents dataset has bee transformed to about linearly separable dataset. Therefore, the classification performance was increased after feature selection process. In order to reduce the complexity of classifier and to increase the performance of classifier algorithms, the feature selection algorithms could be used prior to classifier algorithms. As can be seen from obtained results, the proposed DMS could be used on prediction of real traffic accidents.

## 5. Conclusions

Using the geographical information system is useful in displaying of the accident locations on the map. At the same time, because of the loss of lives and large amount of money, the researches aiming to prevent the traffic accidents are very popular in developing countries. In this situation, it is very important to forecast the probability of the occurrence of accidents according to the weather conditions for implementing proper precautionary measures in traffic safety studies. Although the study has low accuracy values, use of different artificial intelligence methods with the increase of accidental data warehouse will improve the accuracy of the decision support system. To perform this situation, a novel decision making system has been proposed and applied to prediction of a real traffic accidents on the Konya–Afyonkarahisar highway in Turkey. The proposed method comprised two stages comprising of correlation-based feature selection and classification algorithms including SVM and ANN models. The proposed method has obtained better results than that of alone SVM and ANN models and exhibited the applicability to prediction of traffic accidents. So, it is aimed to obtain more reliable results for the system with collecting more accident data, according to the weather and environment conditions as future work.

## Acknowledgements

This study is supported by the Scientific Research Projects of Selcuk University.

## References

Andreescu, M. P., & Frost, D. B. (1998). Weather and traffic accidents in Montreal, Canada. *Climate Research, 9*, 225–230.

Camastra, F. (2007). A SVM-based cursive character recognizer. *Pattern Recognition, 40*(12), 3721–3727.

Comak, E., Arslan, A., & Türkoğlu, İ. (2007). A decision support system based on support vector machines for diagnosis of the heart valve diseases. *Computers in Biology and Medicine, 37*(1), 21–27.

Erdogan, S., Yilmaz, I., Baybura, T., & Gullu, M. (2007). Geographical information systems aided traffic accident analysis system case study: City of Afyonkarahisar. *Accident Analysis and Prevention*.

Haykin, S. (1999). *Neural networks: A comprehensive foundation*. Prentice Hall. ISBN: 0-13-273350-1.

Huang, C.-L., Chen, M.-C., & Wang, C.-J. (2007). *Expert Systems with Applications, 33*(4), 847–856.

Lawrence, J. (1994). *Introduction to neural networks*. California Scientific Software Press. ISBN: 1-883157-00-5.

Levine, N., Kim, K. E., & Lawrence, H. N. (1995). Spatial analysis of Honolulu motor vehicle crashes: I. Spatial patterns. *Accident Analysis and Prevention, 27*(5), 663–674.

Lewis, D., & Gale, W. (1994). A sequential algorithm for training text classifiers. In *Annual ACM conference on research and development in information retrieval, the 17th annual international ACM SIGIR conference on research and development in information retrieval*, New York (pp. 3–12).

Loo, B. P. Y. (2006). Validating crash locations for quantitative spatial analysis: A GIS based approach. *Accident Analysis and Prevention, 38*(1), 879–886.
Ma, Y., Guo, L., & Cukic, B. (2009). A statistical framework for the prediction of fault-proneness. *Advances in machine learning application in software engineering* (pp. 11–12). Idea Group Inc.
Müge, K., & Şenkal, Ş. (1999). Accident black spot determination methods, II. Transportation and traffic congress – Exhibitions book of notifications. Publishment no.: 242, October 2nd.
Polat, K., & Güneş, S. (2007). Breast cancer diagnosis using least square support vector machine. *Digital Signal Processing, 17*(4), 694–701.
http://research.microsoft.com/≃jplatt/svm.html (last arrived 2009).
Sawalha, Z., & Sayed, D. (2006). Transferability of accident prediction models. *Safety Science, 44*, 209–219.
Vapnik, V. (1995). *The nature of statistical learning theory*. New York: Springer.
http://en.wikipedia.org/wiki/Binary_classification (last accessed 2009).
http://www.com/Mathbits/TISection/Statistics2/correlation.htm (last accessed 2009).
www.mathworks.com (last accessed 2009).