Journal of Statistics Applications & Probability
An International Journal

http://dx.doi.org/10.12785/jsap/020303

Large-Scale Statistical Modelling via Machine Learning Classifiers

*Christina Parpoula**, *Krystallenia Drosou** *and Christos Koukouvinos**

Department of Mathematics, National Technical University of Athens, Zografou 15773, Athens, Greece

Received: 29 Mar. 2013, Revised: 25 Jun. 2013, Accepted: 1 Jul. 2013
Published online: 1 Nov. 2013

Abstract: The problem of statistical modelling and identifying the significant variables in large data sets is common nowadays. This paper deals with the statistical analysis of two large dimensional data sets; we firstly conduct a seismic hazard sensitivity analysis using seismic data from Greece acquired during the years 1962 – 2003, and then analyze Trauma data collected in an annual registry conducted during the year 2005 by the Hellenic Trauma and Emergency Surgery Society involving 30 General Hospitals in Greece. The main purpose of both analyses is to extract high-level knowledge for the domain user or decision-maker. Eight non parametric classifiers derived from data mining methods (Multilayer Perceptrons (MLP) Neural Networks, Radial Basis Function Neural (RBFN) Networks, Bayesian Networks, Support Vector Machines (SVMs), Classification and Regression Tree (C&RT), Chi-square Automatic Interaction Detection (CHAID), C5.0 algorithm and Quick, Unbiased, Efficient Statistical Tree (QUEST)) are employed in this work, and are compared to Logistic Regression and ℓ1-norm SVM in terms of overall classification accuracy, sensitivity, specificity, and Area under the ROC curve (AUROC). The goal of this paper is twofold; assess the importance of several input variables in order to detect the possible risk factors of large earthquakes or to prevent trauma deaths, and examine which classifiers are most suited for a large dimensional data analysis, detecting effectively complex nonlinear relationships and potentially lead to more accurate predictions.

Keywords: Artificial Neural Networks, Bayesian Networks, Decision Trees, Large Dimensional Data, Logistic Regression, Model selection, Sensitivity Analysis, Support Vector Machines
AMS Subject Classification: 62-07, 62H30, 62P12, 68T10

1 Introduction

Several variable selection methods have been developed over the last decade to cope with statistical modelling problems for large dimensional data. The main goal in such problems is to select correctly and parsimoniously the features that influence significantly the response variable. In response to such problems, Data Mining (DM) techniques are widely used for extracting nontrivial, implicit and previously unknown information from massive databases, and effectively applying it for decision making or prediction. Large dimensional variable selection problems arise from diverse fields of sciences, such as seismology (see, for example, [32] and [11]), geosciences (see, for example, [2]) and biomedicine (see, for example, [27]).

Koukouvinos et al. [28] have already dealt with the problem of high-dimensional seismic data analysis by employing statistical methods such as the nonconcave penalized likelihood methods and best subset techniques which were then compared to decision tree methods. Classical variable selection methods, statistics and soft computing techniques have been studied extensively to analyze earthquake data (see, for example, [1], [42] and [40]). Bayesian networks [29] and Artificial Neural Networks (ANNs) (see, for example [15], [3] and [30]) have been widely developed for dealing with nonlinear seismic data. Recent research has targeted on novel data mining techniques [47], computational seismic algorithms [34], and new ANN approaches (see, for example, [25], [26] and [12]) for earthquake hazard analysis. Methods of analysis such as Support Vector Machines (SVMs) are still at an early stage [9]. In this study, we compare several DM techniques with SVMs and ANNs, which are particularly suited for our analysis, since both are flexible models that can cope with complex nonlinear relationships [20].

* Corresponding author e-mail: parpoula.ch@gmail.com, drosou.kr@gmail.com, ckoukouv@math.ntua.gr

DM also plays a significant role in large dimensional medical data analysis since modern hospitals are well equipped with data collection devices, and achieve to collect and store huge databases [52]. Currently, many studies have been carried out and focused on DM applications for biomedical data analysis, for example see [6], [14], [10], [37], [38], and [27]. In this study, DM techniques are performed for trauma data analysis, and are compared to SVMs and ANNs in order to generate a predictive model that can be applied for prediction and classification of new data.

In this paper, our aim is twofold, namely, to build a meaningful and parsimonious model by estimating several seismological or medical parameters, and figure out the link between statistics and data mining for analyzing two important real life problems; earthquake prediction and trauma death prediction. However, analyzing real data is often a complex and laborious process because it is difficult or even impossible to establish a physical relationship that connects a collection of inputs to a particular target or outcome [50]. The main goal is to identify those factors that may contribute to a large earthquake or a trauma death, and the exploration of the way they could do this. This can not be accomplished by using traditional variable selection techniques which often neglect some underlying factors. The innovative nature of our study lies on the comparison of several high-powered data mining techniques that enabled us to deal with mass of seismic and medical data and effectively analyze two real life problems. After performing model selection, we compute several classification metrics and graphics, including the sensitivity analysis procedure.

The rest of this paper is organized as follows. In Section 2, we describe briefly the considered classifiers. In Section 3, we perform the classifier-based analyses, and then carry out a comparative study to evaluate the merits of each employed method. We also describe the performance criteria used for the evaluation of the employed methods. Finally, in Section 4, the obtained results are discussed and some concluding remarks are made.

2 Classifiers

2.1 Logistic regression (LR)

A LR model in which the response variable y has only two possible outcomes, denoted by 0 and 1, is considered. We now present the LR model used in our comparative study.

Suppose there are n experimental runs with a binary response. If we arbitrarily write $y=1$ for a success and $y=0$ for a failure, then we are truly modeling the mean response $P(x_i)$, where $P(x_i)$ is the success probability and x_i denotes the set of covariates or regressors at the ith data point. The logistic model for $P(x_i)$ is then given by

$$P(x_i) = 1/(1+e^{-x_i'\beta}),$$

where the term $x_i'\beta = \beta_0+\beta_1 x_{i1}+...+ \beta_m x_{im}$ is the linear link function. For more details on the LR model, we refer the interested reader to [36].

2.2 Decision Trees (DT)

Decision tree algorithms repeatedly split the data set according to a criterion that maximizes the separation of the data, resulting in a tree-like structure which includes only the important attributes that really contribute to the decisions making. This greedy construction process of decision trees gives the opportunity to develop classification models that may be used to predict or classify future data sets, according to a number of provided decision rules. In this study, we focus on four widely used decision tree algorithms, named as Classification and Regression Tree (C&RT) [7], C5.0 [41], Chi-square Automatic Interaction Detection (CHAID) [24], and Quick, Unbiased, Efficient Statistical Tree (QUEST) [31].

2.3 Artificial Neural Networks (ANNs)

Among the emerging scientific methods for data analysis, computational intelligence methods such as ANNs [44] find applications in seismic data analysis. ANNs are widely used in data mining methodology for revealing hidden non-linear relationships among data [49]. Two ANN algorithms are tested in this paper: Multilayer Perceptron (MLP) and Radial-Basis Function Network (RBFN). MLP is a general purpose feedforward network [21] that uses the classical back propagation algorithm based on a deterministic gradient descent algorithm in order to optimize the error function. RBFN is based on a clustering procedure for computing distances among each input vector and the center, represented by the radial unit [33], [23].

The ANN used in this study is a standard three-layered network with an input, a hidden, and an output layer. The input layer consists of 9 input neurons representing the risk factors (predictor variables), the hidden layer consists of 3

hidden neurons, and the output layer consists of one output neuron modelling the dichotomous outcome (valued as 1 for the positive response, and 0 for the negative response). Since there does not exist a commonly accepted theory for predetermining the optimal number of neurons in the hidden layer, we follow a trial-and-error process and several number of units in the hidden layer are examined, such as 3, 5, 7 and 9. The number of neurons in the hidden layer is set to 3 empirically, e.g., by two-fold cross-validation. For both ANNs the network is trained on the training set, and accuracy is estimated based on the test set. For a detailed description of ANNs, the interested reader may refer to [4].

2.4 Bayesian Network Models

A Bayesian network is a parameter learning structure that provides a succinct way of describing the joint probability distribution for a given set of random variables. A Bayesian network is a graphical probability model that displays variables (often referred to as nodes) in a dataset and the probabilistic, or conditional, independencies between them [16]. In this study, we focus on the Tree Augmented Naive Bayes (TAN) network that is primarily used for classification. The TAN creates a simple Bayesian network model that is an improvement over the standard Naive Bayes model. This is because it allows each predictor to depend on another predictor in addition to the target variable, thereby increasing the classification accuracy.

2.5 Support Vector Machines (SVMs)

SVMs are a comparatively new classification technique based on ideas originated in statistical learning theory [46], [8]. SVMs are a supervised learning method that generates input-output mapping functions from a set of training data. Specifically, the training vectors are mapped into a high-dimensional feature space so that data points can be categorized. Since in our study we deal with a binary classification problem, the two groups are separated in a higher-dimension hyper-plane accordingly to a decision function. The optimal, in terms of classification performance, hyper-plane is the one with the maximal margin of separation between the two classes [46]. SVMs use either linear and non-linear kernel functions to transform input data to a high-dimensional feature space in which the input data become more separable compared to the original input space.

Though several kernels such as Laplace, Bessel, Anova, and Spline have been proposed by researchers [22], in this study we examine the following four basic kernels, which are the most widely used and are defined as:

1.Linear: $K(x_i,x_j) = x_i^T x_j$;
2.Polynomial: $K(x_i,x_j) = (\gamma x_i^T x_j + coef0)^d$;
3.Gaussian Radial Basis Function: $K(x_i,x_j) = exp(-\gamma\|x_i - x_j\|^2)$;
4.Sigmoid: $K(x_i,x_j) = \tanh(\gamma x_i^T x_j + coef0)$

where $\gamma > 0$, coef0 and d (degree) are kernel parameters.

We examine different kernel functions in order to obtain the best model, as they each use different algorithms and parameters. The goal is to find the optimum balance between a wide margin and a small number of misclassified data points. Each kernel function has a regularization penalty parameter (known as C) which controls the trade-off between these two values. We need to experiment with different values of this and other kernel parameters in order to find the best model, which means that training a SVM for the above kernels requires the setting of 1, 4, 2, and 3 parameters respectively.

Although SVMs are commonly developed as a method of finding the maximal margin hyperplane, SVMs can also be formulated as a regularized function estimation problem, corresponding to a hinge loss function plus an $\ell 1$-norm or an $\ell 2$-norm regularization term on the fitted coefficients [51]. The Least Absolute Shrinkage and Selection Operator (LASSO) method [45] is a common approach in regression for parameter estimation. Bradley and Mangasarian (1998) [5] adapted the $\ell 1$-regularization (LASSO) to SVM. Zhu et al. (2003) [53] extended the idea of using $\ell 1$-norm constraints for automatical variable selection to classification problems by proposing an $\ell 1$-norm SVM. Fung and Mangasarian (2004) [17] proposed a fast $\ell 1$ SVM modification using a newton linear programming SVM. Recently, Wang et al. (2006) [48] adapted the elastic net penalty term [54] to SVMs by using a mix of the $\ell 1$-norm and the $\ell 2$-norm penalties. The elastic net SVM is especially useful for cases in which the number of the relevant variables exceeds the sample size. The $\ell 1$-norm SVM is suitable for our real data analyses (seismic and trauma), since the dimension of the data is not larger than the number of training samples [51].

3 Comparative Analysis - Results

3.1 Model Evaluation

Assessing the reliability of classifier algorithms is essential to ensure data quality. The most common criterion to assess the quality of a classification model is discrimination which measures how well the two classes in the data set are separated [49]. We consider the four most commonly used measures of discrimination for evaluating the performance of the employed methods. Initially, the accuracy is used as first criterion. Accuracy is defined as the percentage of correct classified records in training, test or validation set for every used method. The other two criteria used are the sensitivity and specificity which are two statistical measures of the performance of a binary classification test and are closely related to the concepts of Type I and Type II errors. Sensitivity measures the proportion of actual positives which are correctly identified as such whereas specificity measures the proportion of negatives which are correctly identified. Another popular statistical tool for describing accuracy is the Receiver Operating Characteristic (ROC) curve [39]. An ROC curve by definition is used to evaluate the performance of a system with dichotomous outcomes. Traditionally, the Area Under the ROC Curve (AUROC) is used as a summary index of test accuracy [19], and is useful as a descriptive of overall test performance.

More precisely, given a classifier and a record, there are four possible scenarios. Positive records are correctly predicted as positive (True Positive-TP), positive records are incorrectly identified as negative (False Negative-FN), negative records are classified as positive ones (False Positive-FP) and finally negative records are correctly identified as negative (True Negative-TN). Using a two-by-two confusion matrix we can easily represent these possible outcomes and compute the four measures, since they are defined as follows:

$$\text{Accuracy (ACC)}= \frac{TP+TN}{TP+FP+TN+FN}$$

$$\text{Sensitivity} = \frac{TP}{TP+FN} = \text{1-Type II error}$$

$$\text{Specificity} = \frac{TN}{TN+FP} = \text{1-Type I error}$$

The ROC curve calculates the Sensitivity= 1-False Negative Rate as a function of (1-Specificity)= False Positive Rate for all the possible cutoffs. In our study, the models' performance is assessed by calculating and comparing the AUROCs.

3.2 Classifier Performance for Seismic Data

In this section, we compare several classifiers from the machine learning field, and then report the results of the modelling process. To provide an unbiased estimate of each model's discrimination, the performance criteria values are calculated from a data set not used in the model building process. Usually, a portion of the original data set, called the test set, is put aside for this purpose [13]. A classifier should present high values of ACC, sensitivity, specificity and AUROC, and the model's generalization performance is often estimated by the holdout validation (i.e., train/test split) [20]. Here, we deal with a high-dimensional data set consisting of 10333 earthquakes, that is split randomly into a training set, containing 75% of cases (7749) and the test set, containing 25% of cases (2584) in order to evaluate the performance of classifiers on new data. For the partitioning, the total observations are randomly selected to create the training and the test sets, according to their predefined size.

The examined data set consists of the response variable y that refers to the earthquake's Magnitude coding as $0,1$ (magnitude$>$ 6.5: 1, otherwise: 0), where the magnitude scale is used to express the seismic energy released by each earthquake, 9 statistically significant factors obtained by the performance of the variable selection techniques in [28], and 10333 instances. The names of these factors are included in the Appendix Section. The analysis was carried out using the SPSS 17.0 and SPSS Clementine 12.0 statistical software. The interested reader may refer to [43] for the MATLAB implementation of the ℓ1-norm SVM.

3.2.1 LR for Seismic Data

The estimated model derived from LR methodology is:

$$1.8814-1.5934*x_1+0.4791*x_2+0.4844*x_4+0.9662*\ x_5+\ 1.2827*x_7\ -0.1536*x_8\ -2.3002*x_9-1.3152*x_{10}+1.3955\ x_{11}$$

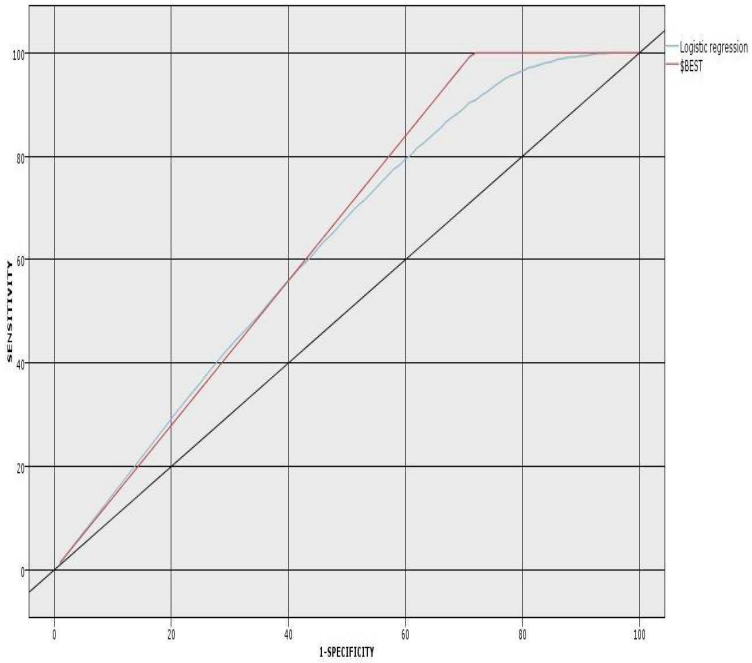

Fig. 1: ROC curve for the estimated LR model (seismic)

Fig. 1 displays the ROC curve for the estimated LR model. The further the curve lies above the reference line, the more accurate the test. The AUROC achieved the high value of 0.90.

All variables are ranked from the 1st to the 9th according to the value of the test statistic (i.e., according to their significance/contribution into the model). We note, that this procedure was applied to all methods, and no differences on the ranking were observed. The final arrangement was obtained from all statistical analysis methods simultaneously. So the final setting to order from the most to the least significant variable is: x_1, x_7, x_{11}, x_5, x_9, x_2, x_{10}, x_4, x_8.

We observe here that the most significant factor for seismological data was found to be the variable "year". Specifically, according to its negative coefficient in the model we conclude that in the last years, earthquakes of lower magnitude were observed. This correlation between "year" and "magnitude", may indicate a periodicity in seismic activity and it needs further investigation from seismologists. This issue is very crucial on earthquake prediction and estimation of earthquake effects on building, and in last decades many researchers have focused on this. The second statistically significant factor is "hyper distance", which is the Euclidean distance, meaning hyper distance $= \sqrt{(epicenterx)^2 + (epicentery)^2}$. These two geographical characteristics of epicenter may be useful for determining areas that present high hazard for strong earthquakes. Seismologists are very interested in studies about such areas since they can combine these ones with their findings about tectonic plates. Taking into consideration these features and the time moments that energy releases through earthquakes, researchers have useful information for earthquake prediction. One more important factor is "depth". The result is presumable since it is well known that the stronger earthquake activity is observed in the external surface of Earth. Among the 9 factors including in the data set, there are factors that are intrinsic to the earthquake, such as the depth ($x11$), geologic characteristics such as, the hyper distance ($x7$) and geographical characteristics, such as the latitude ($x4$), the azimuth ($x8$), the ordinate of the epicenter ($x9$) and the abscissa of epicenter ($x10$).

3.2.2 DT for Seismic Data

C5.0 algorithm has clearly better classification accuracy, sensitivity, and specificity which reaches the absolute percentage of 100% for the training and the test set. Comparing CHAID and C&RT, the first one has more correctly classified records in test set (99.35%) whereas the corresponding percentages for training set are the same for both classifiers. CHAID

Table 1: Advanced comparison of decision trees performance (seismic)

	ACC		SENSITIVITY		SPECIFICITY	
Classifier	Training set	Test set	Training set	Test set	Training set	Test set
C5.0	100%	100%	100%	100%	100%	100%
CHAID	99.24%	99.35%	99.7%	99.72%	98.05%	98.49%
C&RT	99.24%	99.25 %	98.95%	99.14%	100%	100%
QUEST	99.07%	98.85 %	97.64%	97.89%	97.43%	97.95%

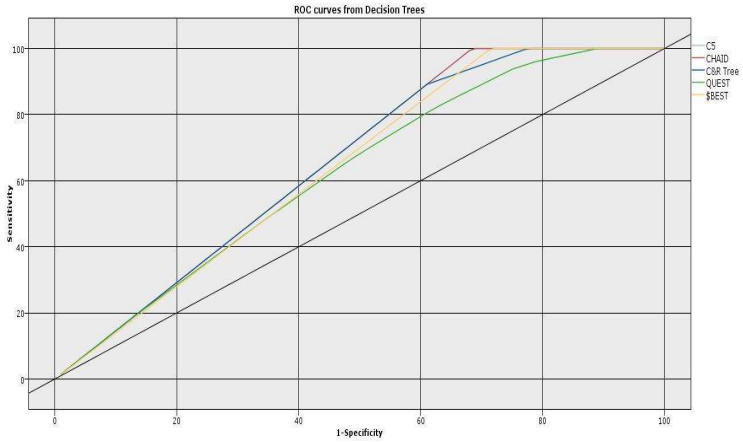

Fig. 2: ROC Curves derived from decision trees (seismic)

has adequate specificity whereas the same measure for C&RT is 100%, which means that the classifier recognizes all actual negatives; in other words this means that C&RT has low Type I error rate. This measure alone does not tell us how well the classifier recognizes positive cases and so it is necessary to take into consideration both sensitivity of the used classifiers. When the two algorithms are evaluated against the sensitivity, CHAID has clear advantage having highest percentages, which means that the Type II error rates are lower than the ones of C&RT algorithm. QUEST has the worst performance compared to C5.0, CHAID and C&RT in terms of ACC, sensitivity and specificity. Fig. 2 displays the ROC curves derived from all decision tree algorithms. The AUROC is 1, 1, 0.98 and 0.95 for the C5.0, CHAID, C&RT and QUEST respectively. In general, C5.0 and CHAID seem to outperform C&RT, and then follows the QUEST with the worst performance criteria values.

3.2.3 ANNs for Seismic Data

Table 2: Estimated Accuracy of ANNs (seismic)

Method	Hidden Units	Estimated Accuracy (%)
RBFN	**3**	**83.51**
RBFN	5	81.23
RBFN	7	80.53
RBFN	9	81.70
MLP	**3**	**97.69**
MLP	5	97.5
MLP	7	97.45
MLP	9	97.33

We examine several number of units such as 3, 5, 7 and 9 in the hidden layer in order to determine the optimal number of neurons in the hidden layer. Table 2 shows the estimated accuracy of binary classification by ANNs with 3, 5, 7 and 9 units in the hidden layer for each MLP and RBFN method. The number of neurons in the hidden layer is set to 3 since this value is found to be optimal resulting in higher estimated accuracy for both MLP and RBFN methods.

MLP network has clearly higher classification accuracy compared to RBFN and Bayesian network for the training and the test set. Furthermore, MLP network achieves excellent results for specificity, and reaches the absolute percentage of

Table 3: Advanced comparison of networks performance (seismic)

	ACC		SENSITIVITY		SPECIFICITY	
Classifier	Training set	Test set	Training set	Test set	Training set	Test set
MLP (Neural Net 1)	97.46%	97.37%	100 %	100 %	91.08%	91.65 %
Bayesian Network	87.65%	88.73 %	89.96 %	91.21 %	81.85 %	83.53 %
RBFN (Neural Net 2)	83.4%	84.11%	84.01 %	84.82 %	81.85%	82.57 %

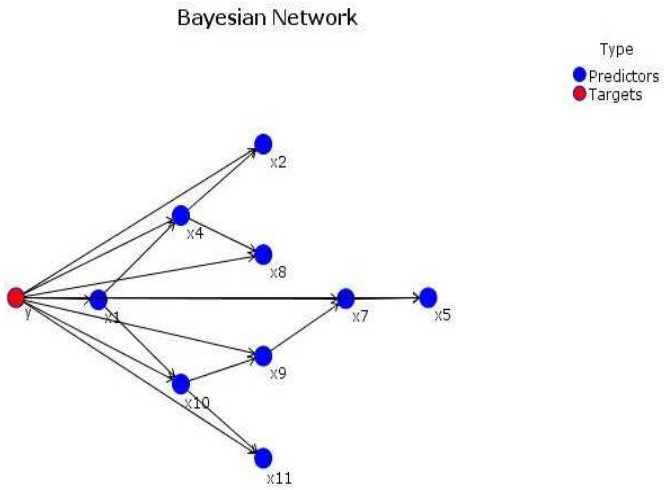

Fig. 3: Bayesian Network (seismic)

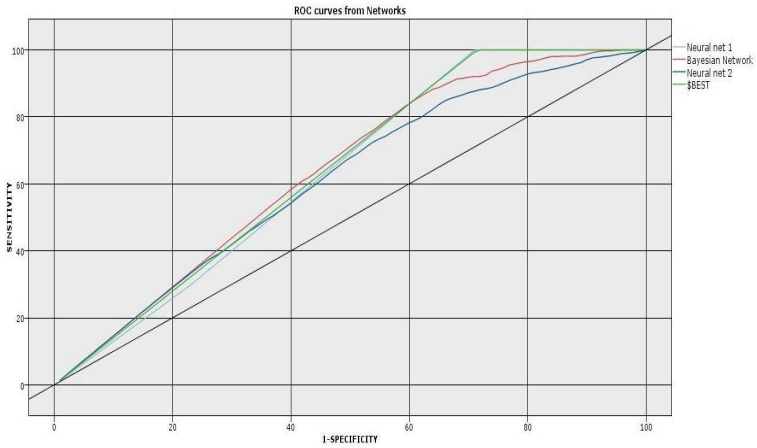

Fig. 4: ROC Curves derived from networks (seismic)

100% for sensitivity. Fig. 3 illustrates the Bayesian network for the 9 statistically significant variables. Fig. 4 displays the ROC curves derived from MLP, RBFN and Bayesian networks. The AUROC is 0.94, 0.93 and 0.86 for the MLP, Bayesian and RBFN network respectively. In general, MLP network technique outperforms Bayesian network, and then follows RBFN taking into consideration all sets resulting from partitioning.

3.2.4 SVMs for Seismic Data

Besides the kernel function, the regularization parameter C and the value of gamma γ are selected from several candidates, and the one that results in the best performance is chosen. If the kernel type is set to polynomial or sigmoid the parameter bias sets the coef0 value in the kernel function and the default value 0 is suitable in most cases. The parameter degree is enabled only if kernel type is set to polynomial and controls the complexity (dimension) of the mapping space (in our study we set the value $d = 3$).

The regularization parameter C controls the trade-off between maximizing the margin and minimizing the training error term. This value should normally be between 1 and 10 inclusive. Increasing the value improves the classification accuracy for the training data, but this can also lead to overfitting. The gamma value should normally be between $3/k$ (=0.333) and $6/k$ (=0.666), where k is the number of input fields (9 in our study).

SVM parameter selection can be viewed as an optimization process, since a grid search method is performed to control parameters C and γ, and obtain the best possible model. Table 4 and 5 show the results of a grid-based search for our data set using the four kernels. In this comparative study, the best model with the highest estimated accuracy is obtained using $C = 10$ and $\gamma = 0.66$ for all considered kernels. After detecting the best regularization parameters, we train our final model and estimate the predictive accuracy.

Table 4: Results of grid search for Gaussian radial basis function (seismic)

Predictive Accuracy (%)										
Gaussian RBF										
0.34	c=1	c=2	c=3	c=4	c=5	c=6	c=7	c=8	c=9	**c=10**
Training Set	86.24	87.33	87.81	88.05	88.31	88.45	88.62	88.69	88.83	**88.94**
Test set	86.44	86.98	87.45	87.76	88.03	88.3	88.53	88.76	88.96	**89.23**
0.35	c=1	c=2	c=3	c=4	c=5	c=6	c=7	c=8	c=9	**c=10**
Training Set	86.51	87.47	87.9	88.24	88.39	88.56	88.66	88.83	89.01	**89.13**
Test set	86.48	87.1	87.6	87.99	88.22	88.53	88.76	89.07	89.35	**89.42**
0.4	c=1	c=2	c=3	c=4	c=5	c=6	c=7	c=8	c=9	**c=10**
Training Set	86.96	87.73	88.13	88.44	88.54	88.85	89.01	89.11	89.25	**89.49**
Test set	86.94	87.52	87.95	88.22	88.8	89.07	89.42	89.42	89.69	**89.89**
0.45	c=1	c=2	c=3	c=4	c=5	c=6	c=7	c=8	c=9	**c=10**
Training Set	87.19	87.95	88.44	88.61	88.89	89.14	89.29	89.51	89.71	**89.89**
Test set	87.18	87.87	88.18	88.88	89.27	89.46	89.73	89.85	89.85	**89.97**
0.5	c=1	c=2	c=3	c=4	c=5	c=6	c=7	c=8	c=9	**c=10**
Training Set	87.45	88.13	88.54	88.91	89.18	89.4	89.64	89.95	90.14	**90.36**
Test set	87.45	87.07	88.61	89.31	89.46	89.73	89.97	89.97	90	**90.31**
0.55	c=1	c=2	c=3	c=4	c=5	c=6	c=7	c=8	c=9	**c=10**
Training Set	87.58	88.45	88.78	89.2	89.46	89.8	90.07	90.29	90.43	**90.56**
Test set	87.64	88.26	89.07	89.46	89.73	90.04	90	90.43	90.59	**90.78**
0.6	c=1	c=2	c=3	c=4	c=5	c=6	c=7	c=8	c=9	**c=10**
Training Set	87.9	88.51	89.28	89.37	89.85	90.07	90.36	90.42	90.66	**90.94**
Test set	87.72	88.49	89.35	89.66	89.97	90.12	90.47	90.66	90.78	**91.24**
0.65	c=1	c=2	c=3	c=4	c=5	c=6	c=7	c=8	c=9	**c=10**
Training Set	88.02	88.71	89.22	89.76	90.03	90.38	90.54	90.85	91.02	**91.15**
Test set	87.87	89	89.5	89.93	90.2	90.55	90.78	91.2	91.44	**91.67**
0.66	c=1	c=2	c=3	c=4	c=5	c=6	c=7	c=8	c=9	**c=10**
Training Set	87.96	88.73	89.25	89.81	90.09	90.42	90.6	90.89	91.01	**91.24**
Test set	87.87	88.96	89.54	89.97	90.39	90.59	90.78	91.28	91.44	**91.82**

Table 5: Results of grid search for sigmoid, linear and polynomial kernel (seismic)

Predictive Accuracy (%)										
Sigmoid	c=1	c=2	c=3	c=4	c=5	c=6	c=7	c=8	c=9	**c=10**
Training Set	71.7	71.71	71.71	71.72	71.72	71.72	71.72	71.71	71.71	**71.74**
Test set	68.38	68.46	68.46	68.5	68.5	68.46	68.46	68.46	68.46	**68.5**
Linear	c=1	c=2	c=3	c=4	c=5	c=6	c=7	c=8	c=9	**c=10**
Training Set	83.76	83.93	84	84.08	84.15	84.13	84.13	84.15	84.13	**84.18**
Test set	83.92	84.23	84.35	84.54	84.54	84.62	84.62	84.66	84.7	**84.73**
Polynomial	c=1	c=2	c=3	c=4	c=5	c=6	c=7	c=8	c=9	**c=10**
Training Set	88.34	89.23	90.26	90.96	91.34	91.65	91.96	92.13	92.52	**92.87**
Test set	88.45	89.11	90.55	91.13	91.71	91.98	92.13	92.44	92.75	**93.26**

The SVM with a polynomial kernel has clearly higher classification accuracy compared to the ℓ1-norm SVM and the SVMs with RBF, linear and sigmoid kernel for the training and the test set. Furthermore, the polynomial SVM achieved the highest values for sensitivity and specificity. The Gaussian SVM (RBF) outperforms the linear SVM and the ℓ1-norm SVM in terms of ACC, sensitivity and specificity, and then follows the sigmoid SVM. Fig. 5 displays the ROC curves derived from all SVMs with the four considered kernels, and Fig. 6 displays the ℓ1-norm SVM. The AUROC is 0.98, 0.97, 0.90, 0.62 for the SVMs with a polynomial, RBF, linear and sigmoid kernel respectively. The AUROC for the ℓ1-norm SVM takes the lowest value equal to 0.59. The ℓ1-norm SVM performs almost similarly to the linear SVM in terms of ACC, sensitivity and specificity. Note here, that the ℓ1-norm SVM identifies all 11 possible risk factors of large

Table 6: Advanced comparison of SVMs performance (seismic)

	ACC		SENSITIVITY		SPECIFICITY	
Classifier	Training set	Test set	Training set	Test set	Training set	Test set
Polynomial (SVM 2)	92.86%	93.52%	95.05 %	95.97 %	87.37%	87.36 %
Gaussian RBF (SVM 1)	91.24%	91.82 %	93.82 %	94.79 %	84.75 %	85.39 %
Linear (SVM 4)	84.35%	84.73%	90.76 %	92.52 %	67.51%	67.85 %
ℓ1-norm SVM	83.96%	84%	89.9%	90%	69.62%	70%
Sigmoid (SVM 3)	71.71%	68.46%	99.4 %	99.43 %	23%	14.7 %

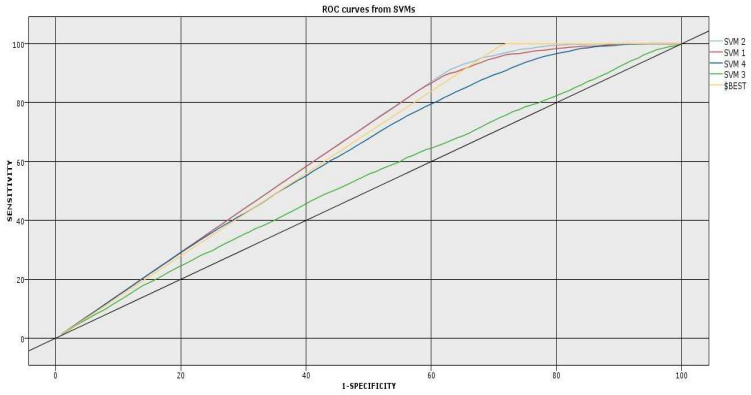

Fig. 5: ROC Curves derived from SVMs (seismic)

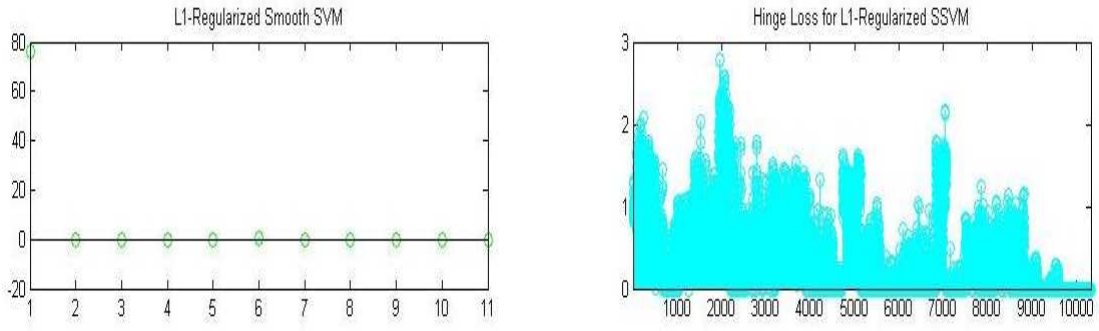

Fig. 6: ℓ1-norm SVM (seismic)

earthquakes as statistically significant. Table 6 shows that the ℓ1-norm SVM tends to have higher Type I errors and lower Type II errors, in other words this means that the ℓ1-norm SVM tends to declare at a higher rate inactive variables to be active, and at a lower rate active variables to be inactive.

3.2.5 Overall Comparison for Seismic Data

Table 7 ranks the best candidate models according to the specified performance criteria, and helps the experimenter to choose the best approach for a given analysis. Fig. 7 displays the ROC curves derived from all methods employed in this comparative study. The further the curve lies above the reference line, the more accurate the test.

Table 7: Advanced comparison of classifiers via overall accuracy and AUROC (presented in descending order)

	Overall Accuracy %		Area Under the ROC curve
Classifier	Training set	Test set	Test set
C5.0	100%	100%	1
CHAID	99.24%	99.35%	1
C&RT	99.24%	99.25 %	0.98
QUEST	99.07%	98.85 %	0.95
Neural Net 1 (MLP)	97.46%	97.37%	0.94
SVM 2 (polynomial)	92.86 %	93.52%	0.98
SVM 1 (RBF)	91.24%	91.82%	0.97
Bayesian Network	87.65 %	88.73%	0.93
SVM 4 (linear)	84.35 %	84.73%	0.90
Logistic Regression	84.13 %	84.23%	0.90
Neural Net 2 (RBFN)	83.4%	84.11%	0.86
SVM 3 (sigmoid)	71.71 %	68.46%	0.62
ℓ1-norm SVM	83.96%	84%	0.59

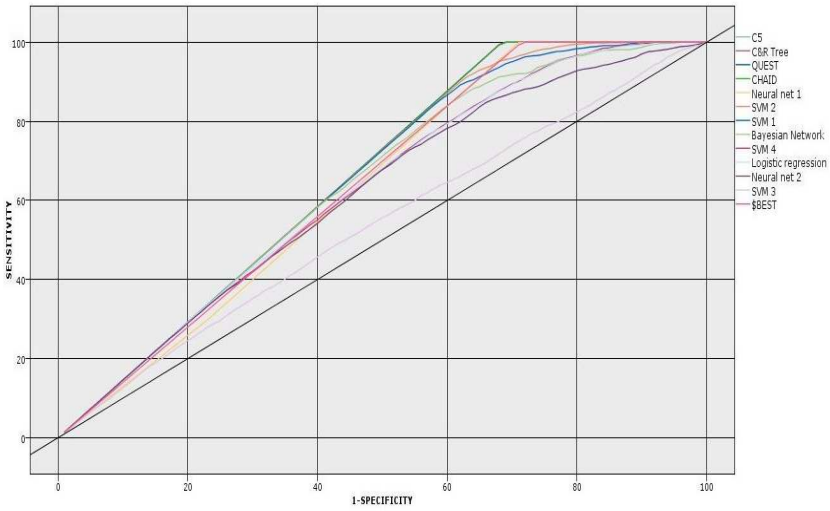

Fig. 7: ROC Curves derived from all models (seismic)

3.3 Classifier Performance for Trauma Data

In this section, we deal with a large dimensional Trauma data set consisting of $N = 8862$ patients and 41 factors that include demographic, transport and intrahospital data used to detect possible risk factors of death. According to medical advices, all the prognostic factors should be treated equally during the statistical analysis and there is no factor that should be always maintained in the model. The data set is split randomly into a training set, containing 75% of cases (6646) and the test set, containing 25% of cases (2216) in order to evaluate the performance of classifiers on new data. For the partitioning, the total observations are randomly selected to create the training and the test sets, according to their predefined size.

For each patient the target attribute, i.e., the probability of death, is reported. This response variable y is binary, expressed in the form of two categories, where 0 value denotes the survival, while the value of 1 denotes the death. The names of these factors are included in the Appendix Section. The analysis was carried out using the SPSS 17.0 and SPSS Clementine 12.0 statistical software. The interested reader may refer to [43] for the MATLAB implementation of the ℓ1-norm SVM.

3.3.1 LR for Trauma Data

The estimated model derived from LR methodology is:

$$-5.97+0.72*x_2+0.25*x_{11}+0.09*x_{16}+0.56*x_{20}+ 0.06*x_{23}+1.02*x_{25}-0.16*x_{27}+1.46*x_{71}+1.30*x_{101}$$

Fig. 8 displays the ROC curve for the estimated LR model. The further the curve lies above the reference line, the more accurate the test. The AUROC achieved the high value of 0.986.

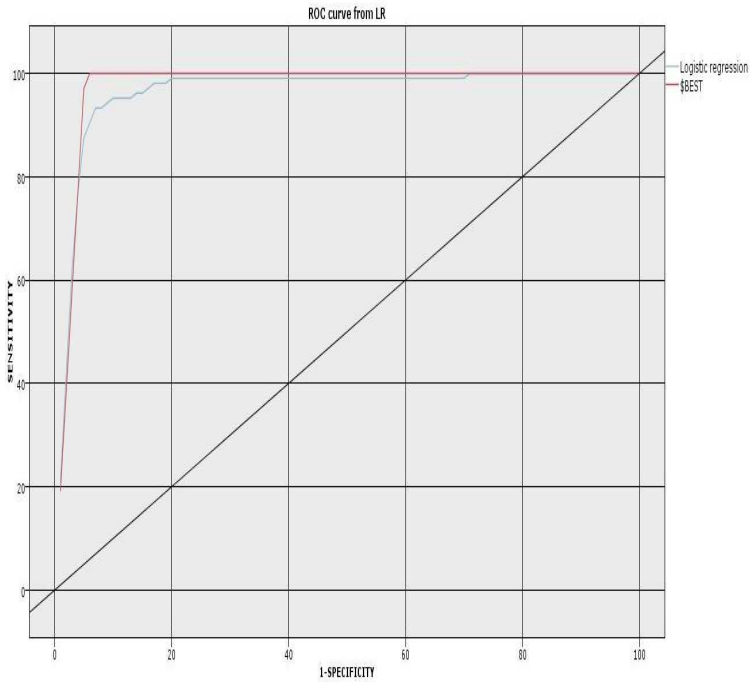

Fig. 8: ROC curve for the estimated LR model (trauma)

All variables are ranked from the 1st to the 9th according to the value of the test statistic (i.e., according to their significance/contribution into the model). The final arrangement was obtained from all statistical analysis methods simultaneously. So the final setting to order from the most to the least significant variable is: x_{71}, x_{20}, x_{101}, x_{25}, x_{11}, x_{23}, x_{16}, x_{27}, x_2.

3.3.2 DT for Trauma Data

Table 8: Advanced comparison of decision trees performance (trauma)

	ACC		SENSITIVITY		SPECIFICITY	
Classifier	Training set	Test set	Training set	Test set	Training set	Test set
C5.0	99.05%	98.96%	84.21 %	82.69 %	99.85 %	99.76 %
CHAID	98.38%	98.19 %	80.12 %	78.90%	99.36%	99.09 %
C&RT	98.54%	98.64 %	77.19 %	78.84%	99.69%	99.62 %
QUEST	98.36%	98.64 %	76.60 %	78.76%	99.54 %	99.52 %

C5.0 algorithm has clearly better classification accuracy, sensitivity, and specificity which reaches the percentage of 99% in some cases for the training and the test set. Comparing CHAID and C&RT, the second one has more correctly classified records in both training and test set, and higher specificity which means that the classifier recognizes more actual negatives; in other words this means that C&RT has lower Type I error rate. This measure alone does not tell us how well the classifier recognizes positive cases and so it is necessary to take into consideration both sensitivity of the used classifiers. When the two algorithms are evaluated against the sensitivity, CHAID has clear advantage having highest percentages, which means that the Type II error rates are lower than the ones of C&RT algorithm. QUEST has the worst performance compared to C5.0, CHAID and C&RT in terms of ACC and sensitivity. QUEST has only higher percentages of specificity compared to CHAID. Fig. 9 displays the ROC curves derived from all decision tree algorithms. The AUROC

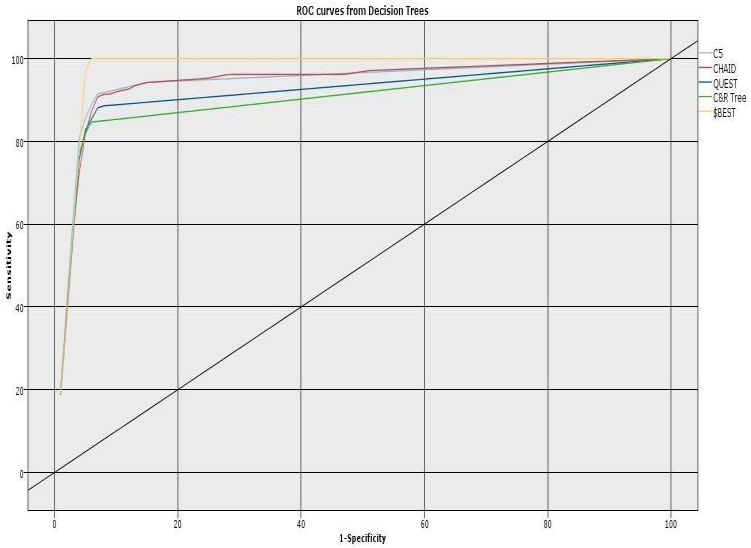

Fig. 9: ROC Curves derived from decision trees (trauma)

is 0.964, 0.964, 0.919 and 0.937 for the C5.0, CHAID, C&RT and QUEST respectively. In general, C5.0 and CHAID seem to outperform C&RT, and then follows the QUEST with the worst performance criteria values.

3.3.3 ANNs for Trauma Data

Table 9: Estimated Accuracy of ANNs (trauma)

Method	Hidden Units	Estimated Accuracy (%)
RBFN	3	94.72
RBFN	5	94.95
RBFN	**7**	**95.07**
RBFN	9	94.99
MLP	3	98.65
MLP	**5**	**98.76**
MLP	7	98.72
MLP	9	98.75

Table 10: Advanced comparison of networks performance (trauma)

	ACC		SENSITIVITY		SPECIFICITY	
Classifier	Training set	Test set	Training set	Test set	Training set	Test set
MLP (Neural Net 1)	98.92%	98.92%	100 %	100 %	100 %	100 %
Bayesian Network	98.38%	96.93 %	91.81 %	85.41 %	98.73 %	98.42 %
RBFN (Neural Net 2)	94.86%	95.3 %	83.04%	83.65 %	99.77 %	99.66 %

We examine several number of units such as 3, 5, 7 and 9 in the hidden layer in order to determine the optimal number of neurons in the hidden layer. Table 9 shows the estimated accuracy of binary classification by ANNs with 3, 5, 7 and 9 units in the hidden layer for each MLP and RBFN method. The number of neurons in the hidden layer is set to 7 for RBFN and 5 for MLP since this value is found to be optimal resulting in higher estimated accuracy for MLP and RBFN methods respectively.

MLP network has clearly higher classification accuracy compared to RBFN and Bayesian network for the training and the test set. Furthermore, MLP network achieves excellent results for sensitivity and specificity, and reaches the absolute

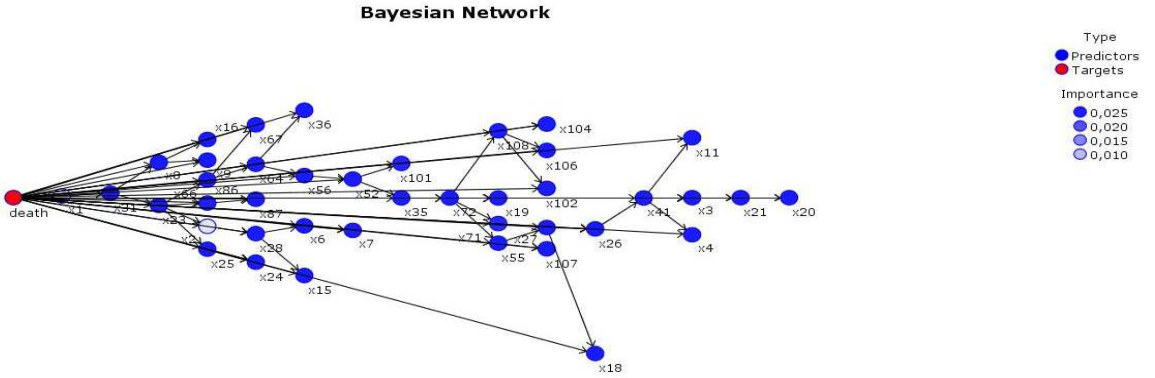

Fig. 10: Bayesian Network (trauma)

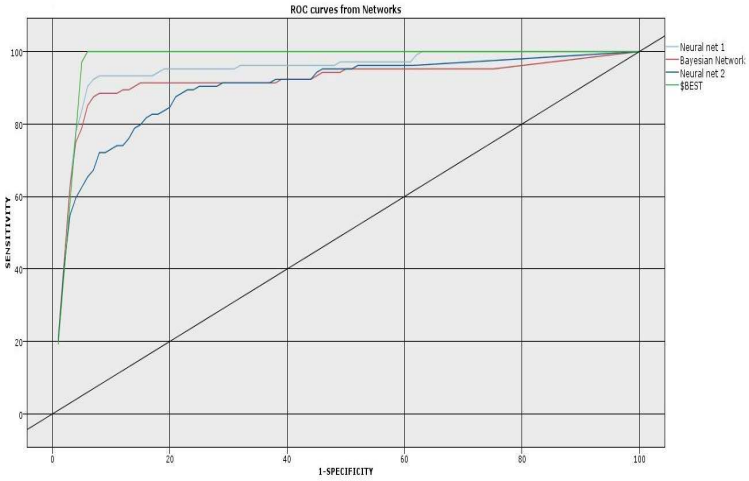

Fig. 11: ROC Curves derived from networks (trauma)

percentage of 100% for both training and test set. The Bayesian network has better performance compared to RBFN in terms of ACC and sensitivity whereas it has lower percentages of specificity for both training and test set. Fig. 9 illustrates the Bayesian network for the 41 input variables. Fig. 10 displays the ROC curves derived from MLP, RBFN and Bayesian networks. The AUROC is 0.98, 0.937 and 0.917 for the MLP, Bayesian and RBFN network respectively. In general, MLP network technique outperforms Bayesian network, and then follows RBFN taking into consideration all sets resulting from partitioning.

3.3.4 SVMs for Trauma Data

Besides the kernel function, the regularization parameter C and the value of gamma γ are selected from several candidates, and the one that results in the best performance is chosen. The gamma value should normally be between $3/k$ (=0.07317) and $6/k$ (=0.14634), where k is the number of input fields (41 in our trauma study). Table 11 and 12 show the results

of a grid-based search for our data set using the four kernels. In this comparative study, the best model with the highest estimated accuracy is obtained using $C = 2$ for the linear kernel, $C = 1$ and $\gamma = 0.146$ for the sigmoid kernel, $C = 6$ or 7 or 8 or 9 or 10 and $\gamma = 0.146$ for the polynomial kernel, and $C = 10$ and $\gamma = 0.146$ for the Gaussian RBF kernel. If the kernel type is set to polynomial or sigmoid the parameter bias sets the coef0 value in the kernel function and the default value 0 is suitable in most cases. The parameter degree is enabled only if kernel type is set to polynomial and is set to be $d = 3$. After detecting the best regularization parameters, we train our final model and estimate the predictive accuracy.

Table 11: Results of grid search for Gaussian radial basis function (trauma)

Predictive Accuracy (%)										
Gaussian RBF										
0.073	c=1	c=2	c=3	c=4	c=5	c=6	c=7	c=8	c=9	**c=10**
Training Set	98.83	99.08	99.13	99.19	99.28	99.32	99.38	99.41	99.46	**99.46**
Test set	98.42	98.42	98.51	98.64	98.64	98.64	98.69	98.69	98.74	**98.74**
0.08	c=1	c=2	c=3	c=4	c=5	c=6	c=7	c=8	c=9	**c=10**
Training Set	98.98	99.11	99.19	99.25	99.32	99.38	99.43	99.46	99.47	**99.52**
Test set	98.42	98.46	98.46	98.55	98.6	98.64	98.64	98.69	98.74	**98.74**
0.09	c=1	c=2	c=3	c=4	c=5	c=6	c=7	c=8	c=9	**c=10**
Training Set	98.99	99.14	99.23	99.32	99.38	99.44	99.47	99.52	99.59	**99.65**
Test set	98.46	98.51	98.51	98.51	98.6	98.64	98.64	98.69	98.69	**98.74**
0.1	c=1	c=2	c=3	c=4	c=5	c=6	c=7	c=8	c=9	**c=10**
Training Set	99.01	99.2	99.28	99.4	99.46	99.47	99.56	99.65	99.67	**99.68**
Test set	98.42	98.51	98.55	98.6	98.6	98.64	98.69	98.74	98.78	**98.78**
0.11	c=1	c=2	c=3	c=4	c=5	c=6	c=7	c=8	c=9	**c=10**
Training Set	99.05	99.26	99.32	99.46	99.46	99.61	99.65	99.68	99.68	**99.71**
Test set	98.51	98.55	98.6	98.64	98.64	98.69	98.74	98.76	98.78	**98.78**
0.12	c=1	c=2	c=3	c=4	c=5	c=6	c=7	c=8	c=9	**c=10**
Training Set	99.05	99.29	99.37	99.46	99.56	99.65	99.68	99.71	99.73	**99.74**
Test set	98.6	98.6	98.64	98.64	98.69	98.74	98.74	98.74	98.78	**98.83**
0.13	c=1	c=2	c=3	c=4	c=5	c=6	c=7	c=8	c=9	**c=10**
Training Set	99.1	99.32	99.44	99.53	99.65	99.68	99.71	99.74	99.76	**99.79**
Test set	98.6	98.64	98.64	98.74	98.74	98.74	98.78	98.78	98.78	**98.83**
0.14	c=1	c=2	c=3	c=4	c=5	c=6	c=7	c=8	c=9	**c=10**
Training Set	99.14	99.34	99.44	99.56	99.68	99.71	99.73	99.77	99.79	**99.85**
Test set	98.6	98.69	98.74	98.74	98.74	98.74	98.78	98.83	98.83	**98.87**
0.146	c=1	c=2	c=3	c=4	c=5	c=6	c=7	c=8	c=9	**c=10**
Training Set	99.19	99.37	99.5	99.64	99.71	99.71	99.74	99.79	99.85	**99.85**
Test set	98.69	98.74	98.74	98.74	98.74	98.74	98.78	98.78	98.83	**98.87**

Table 12: Results of grid search for sigmoid, linear and polynomial kernel (trauma)

Predictive Accuracy (%)										
Sigmoid	**c=1**	c=2	c=3	c=4	c=5	c=6	c=7	c=8	c=9	c=10
Training Set	**95.08**	94.78	94.77	94.78	94.74	94.72	94.75	94.75	94.69	94.66
Test set	**95.35**	95.03	95.03	95.03	94.94	94.94	94.94	94.94	94.94	94.94
Linear	c=1	**c=2**	c=3	c=4	c=5	c=6	c=7	c=8	c=9	c=10
Training Set	98.9	**98.98**	98.95	98.95	98.92	98.92	98.89	98.9	98.9	98.9
Test set	98.69	**98.74**	98.74	98.74	98.69	98.64	98.64	98.64	98.64	98.64
Polynomial	c=1	c=2	c=3	c=4	c=5	**c=6**	**c=7**	**c=8**	**c=9**	**c=10**
Training Set	99.56	99.71	99.79	99.88	99.91	**99.94**	**99.94**	**99.94**	**99.94**	**99.94**
Test set	98.55	98.24	98.37	98.42	98.51	**98.55**	**98.55**	**98.55**	**98.55**	**98.55**

Table 13: Advanced comparison of SVMs performance (trauma)

	ACC		SENSITIVITY		SPECIFICITY	
Classifier	Training set	Test set	Training set	Test set	Training set	Test set
Polynomial (SVM 2)	99.94%	98.55 %	100 %	89.69 %	99.98%	99.56 %
Gaussian RBF (SVM 1)	99.85%	98.51 %	97.66 %	89.58 %	99.96 %	99.52 %
Linear (SVM 4)	98.98%	98.74%	90 %	85.57 %	99.80 %	99.14 %
ℓ1-norm SVM	98%	99%	81.11%	81%	99.53%	100%
Sigmoid (SVM 3)	95.08%	95.35%	83.62 %	85 %	99.65 %	99.28 %

The SVM with a polynomial kernel has clearly higher classification accuracy compared to the ℓ1-norm SVM and the SVMs with RBF, linear and sigmoid kernel for the training and the test set. Furthermore, the polynomial SVM achieved the highest values for sensitivity and specificity. The Gaussian SVM (RBF) outperforms the linear SVM and the ℓ1-norm SVM in terms of ACC, sensitivity and specificity, and then follows the sigmoid SVM. Fig. 12 displays the ROC curves derived from all SVMs with the four considered kernels, and Fig. 13 displays the ℓ1-norm SVM. The AUROC is 0.976, 0.99, 0.98 and 0.858 for the SVMs with a polynomial, RBF, linear and sigmoid kernel respectively. The AUROC for the ℓ1-norm SVM takes the lowest value equal to 0.627. The ℓ1-norm SVM performs almost similarly to the linear SVM in terms of ACC, sensitivity and specificity. The ℓ1-norm SVM detected a set of 39 out of 41 variables as statistically

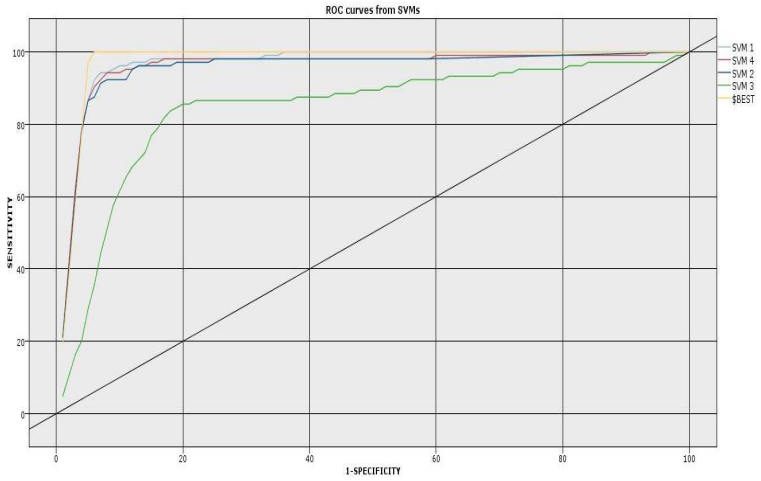

Fig. 12: ROC Curves derived from SVMs (trauma)

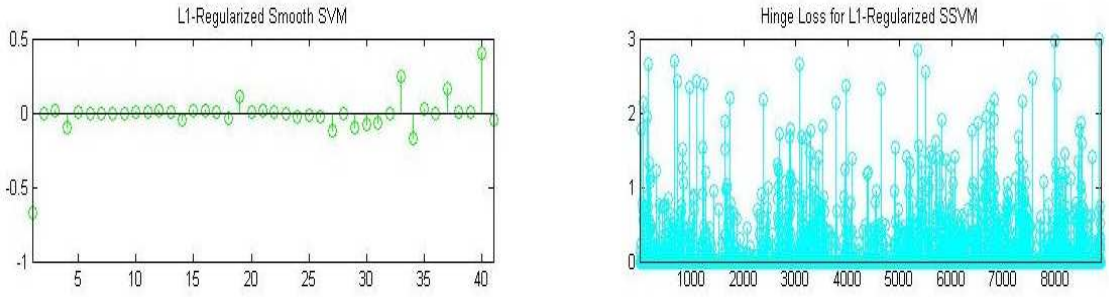

Fig. 13: ℓ1-norm SVM (trauma)

significant which includes the important variables x_{71}, x_{20}, x_{101}, x_{25}, x_{11}, x_{23}, x_{16}, x_{27}, x_2 previously obtained from all statistical analysis methods. The generated model derived from the ℓ1-norm SVM excluded two variables as unimportant, i.e., x_{35} and x_{56}. Note here, that the ℓ1-norm SVM can recognize all actual negatives and reaches the absolute percentage of 100% for specificity, but also has high percentages for sensitivity. Table 13 shows that the ℓ1-norm SVM tends to have lower Type I errors and higher Type II errors, in other words this means that the ℓ1-norm SVM tends to declare at a lower rate inactive variables to be active, and at a higher rate active variables to be inactive.

3.3.5 Overall Comparison for Trauma Data

Table 14 ranks the best candidate models according to the specified performance criteria, and helps the experimenter to choose the best approach for a given analysis. Fig. 14 displays the ROC curves derived from all methods employed in this comparative trauma study. The further the curve lies above the reference line, the more accurate the test.

Table 14: Advanced comparison of classifiers via overall accuracy and AUROC (presented in descending order)

	Overall Accuracy %		Area Under the ROC curve
Classifier	Training set	Test set	Test set
SVM 1 (RBF)	99.85%	98.51%	0.99
Logistic Regression	98.95 %	98.78%	0.986
Neural Net 1 (MLP)	98.95%	98.92%	0.98
SVM 4 (linear)	98.98 %	98.74%	0.98
SVM 2 (polynomial)	99.94 %	98.55%	0.976
C5	99.05 %	98.96%	0.964
CHAID	98.38%	98.19%	0.964
QUEST	98.36%	98.64 %	0.937
Bayesian Network	98.38%	96.93%	0.937
C&RT	98.54%	98.64 %	0.919
Neural Net 2 (RBFN)	94.86%	95.3%	0.917
SVM 3 (sigmoid)	95.08 %	95.35%	0.858
ℓ1-norm SVM	98%	99%	0.627

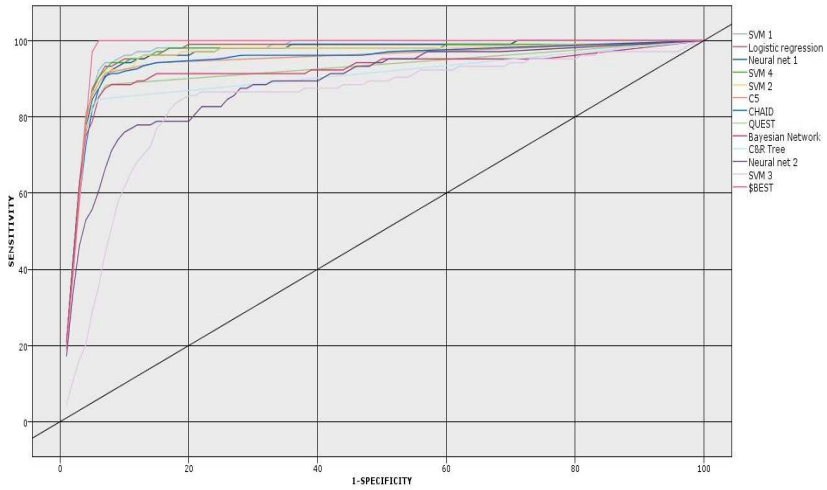

Fig. 14: ROC Curves derived from all models (trauma)

4 Discussion

Recent proliferation of large dimensional databases makes variable selection crucial in model building for large databases with complicated structure. This paper presents an extensive comparative analysis of several machine learning classifiers on real seismological and medical data. C5.0 algorithm had the most excellent results in terms of accuracy, sensitivity and specificity in both studies. Both CHAID and C&RT algorithms, presented very adequate results in these statistical measures about classifiers performance, and outperformed QUEST in both seismic and trauma study. The MLP neural network outperformed greatly the Bayesian network, and then RBFN network followed. The SVM with a polynomial kernel had clearly better classification performance compared to the ℓ1-norm SVM and the SVMs with an RBF, linear and sigmoid kernel in both seismic and trauma study. In general, SVMs (with a polynomial, Gaussian RBF and linear kernel) have been proven for excellent classification performance, since they were found to be more effective than LR method and RBFN network. The RBFN network and the sigmoid SVM were observed to have the worst classification performance for both seismic and trauma study. The ℓ1-norm SVM performed almost similarly to the linear SVM in terms of ACC, sensitivity and specificity in both studies. The ℓ1-norm SVM did not tend to declare at a similar rate inactive variables to be active or active variables to be inactive, hence it could not be considered conservative in this sense.

Assessing the reliability of classifier algorithms is essential to ensure data quality. We used the measures of sensitivity and specificity for the comparison of algorithms in order to provide useful results, since it is obvious that in seismologists effort for earthquakes prediction or in health care domain effort for death prevention of trauma patients a huge problem is arising, obligating them to be more careful in their research. In our seismic study, this obstacle is the twofold hazard of incorrectly earthquake prediction. On one hand, if seismologists result that an earthquake may occur in specific place and time, government should embark on a major earthquake preparedness campaign which costs and frighten citizens. So, if this prediction does not come true, this will have great impact on economy and social life. On the other hand, if a strong earthquake happens without prior information, this will be dangerous and result in many deaths. The value of

this comparative study stands not only in the knowledge discovery, but also in the ability to calculate Type I and Type II error rates for each employed method. In general, we observed that machine learning classifiers identified effectively the most statistically important factors for model building, giving extra significance to the year of seismic activity. Due to this result, seismologists have extra information about periodicity of earthquakes, which is one of key factors for earthquake prediction. In our trauma study, sensitivity and specificity measure the prognostic model's ability to recognize the patients of a certain group (survivors or non-survivors). Identifying an interpretable prognostic model allows medical community to discover previously unknown variable relationships and explore the possible risk factors of death. In general, we observed that machine learning classifiers identified effectively the optimal prognostic model and detected the most statistically significant predictor variables which may assist as guidelines for improving the quality of treatment and therefore survivability of a patient through optimal trauma management.

Neither ANNs nor SVMs are perfect. SVMs are fast in training, but require an appropriate choice of kernel function. ANNs are slower in training, but are fast in classifying and robust to noise. ANNs have been widely developed for dealing with nonlinear seismic or medical data. We hope this work will convince experimenters to use not only ANNs but also SVMs for the extraction of large data set patterns in risk factors of an earthquake or a trauma death. SVMs should be considered a powerful predictive tool to be added to standard LR methodology. Thus, one of the most promising topics for further study is the use of the Support Vector Machines classification technique as an alternative method for supporting seismic and medical knowledge discovery. However, some interesting points are still open and should be investigated in the future. We are currently looking into construction problems of a search method that will automatically identify the best kernel and its parameter settings.

Acknowledgments

The research of the first author was financially supported by a scholarship awarded by the Secretariat of the Research Committee of National Technical University of Athens. The authors would like to thank the referees for their constructive and useful suggestions which resulted in improving the quality of this manuscript.

Appendix

Seismic Data

Y: magnitude($0(< 6.5)$, $1(>= 6.5)$)

–Continuous covariates:
x_1: year, years
x_4: lats, (latitude)
x_7: hyper distance, measured in degrees (^{o})
x_8: azimuth, measured in degrees (^{o})
x_9: epicenterx, the ordinate of epicenter
x_{10}: epicentery, the abscissa of epicenter
x_{11}: depth, the earthquake depth range of 0 - 700 km

–Categorical covariates:
x_2: nome, (1 to 54, all prefectures of Greece)
x_5: intensity, (1 to 12 grades)

Trauma Data

Y: 0 (survival), 1 (death)

–Continuous covariates:
$x1$: weight, kg
$x2$: age, years
$x3$: Glasgow Coma Score, score
$x4$: pulse, N/min
$x6$: systolic arterial blood pressure, mmHg
$x7$: diastolic arterial blood pressure, mmHg
$x8$: Hematocrit (Ht), %
$x9$: haemoglobin (Hb), g/dl
$x11$: white cell count, /ml

$x15$: glucose, mg %
$x16$: creatinine, mg %
$x18$: amylase, score
$x20$: Injury Severity Score, score
$x21$: Revised Trauma Score, score

–Categorical covariates:
$x19$: evaluation of disability (0 = expected permanent big, 1 = expected permanent small, 2 = expected impermanent big,
3 = expected impermanent small, 4 = recovery)
$x23$: cause of injury (0 = fall, 1 = trochee accident, 2 = athletic, 3 = industrial, 4 = crime, 5 = other)
$x24$: means of transportation (0 = airplane, 1 = ambulance, 2 = car, 4 = on foot)
$x25$: Ambulance (0 = no, 1 = yes)
$x26$: hospital of records
$x27$: substructure of hospital (0 = orthopaedic, 1 = CT, 2 = vascular surgeon, 3 = neurosurgeon, 4 = Intensive Care Unit)
$x28$: comorbidities (0 = no, 1 = yes)
$x31$: sex (0 = female, 1 = male)
$x35$: doctor's speciality (0 = angiochirurgeon, 1 = non specialist, 2 = general doctor 3 = general surgeon, 4 = jawbonesurgeon,
5 = gynaecologist, 6 = thoraxsurgeon, 7 = neurosurgeon, 8 = orthopaedic, 9 = urologist, 10 = paediatrician, 11 = children surgeon, 12 = plastic surgeon)
$x36$: major doctor (0 = no, 1 = yes)
$x41$: dysphoria (0 = no, 1 = yes)
$x52$: collar (0 = no, 1 = yes)
$x55$: immobility of limbs (0 = no, 1 = yes)
$x56$: fluids (0 = no, 1 = yes)
$x64$: Radiograph E.R. (0 = no, 1 = yes)
$x66$: US (0 = no, 1 = yes)
$x67$: urea test (0 = no, 1 = yes)
$x71$: destination after the emergency room (0 = other hospital, 1 = clinic, 2 = unit of high care, 3 = intensive care unit I.C.U, 4 = mortuary, 5 = operating room)
$x72$: surgical intervention (0 = no, 1 = yes)
$x86$: arrival at emergency room (0 = 00:00-04:00, 1 = 04:01-08:00, 2 = 08:01-12:00, 3 = 12:01-16:00, 4 = 16:01-18:00, 5 = 18:01-20:00, 6 = 20:01-24:00)
$x87$: exit from emergency room (0 = 00:00-04:00, 1 = 04:01-08:00, 2 = 08:01-12:00, 3 = 12:01-16:00, 4 = 16:01-18:00, 5 = 18:01-20:00, 6 = 20:01-24:00)
$x101$: head injury (0 = none, 1 =$AIS \leqslant 2$,2 =$AIS > 2$)
$x102$: face injury (0 = none, 1 =$AIS \leqslant 2$,2 =$AIS > 2$)
$x104$: breast injury (0 = none, 1 =$AIS \leqslant 2$,2 =$AIS > 2$)
$x106$: spinal column injury (0 = none, 1 =$AIS \leqslant 2$,2 =$AIS > 2$)
$x107$: upper limbs injury (0 = none, 1 =$AIS \leqslant 2$,2 =$AIS > 2$)
$x108$: lower limbs injury (0 = none, 1 =$AIS \leqslant 2$,2 =$AIS > 2$)

References

[1] Ashida, Y., Data processing of reflection seismic data by use of neural network, *Journal of Applied Geophysics*, **35**, 89-98 (1996).
[2] Bao F., He X. and Zhao, F., Applying Data Mining to the Geosciences Data, *2010 International Conference on Computer, Mechatronics, Control and Electronic Engineering (CMCE)*, **5**, 290-293 (2010).
[3] Benbrahim, M., Daoudi A., Benjelloun D. and Ibenbrahim, A., Discrimination of Seismic Signals Using Artificial Neural Networks, *Engineering and Technology*, **4**, 4-7 (2005).
[4] Bishop, C., *Neural Networks for Pattern Recognition*, Oxford, Oxford University Press, (1995).
[5] Bradley, P. S., and Mangasarian O. L., Feature Selection via Concave Minimization and Support Vector Machines, In Shavlik, J. (ed.) *Machine Learning Proceedings of the Fifteenth International Conference(ICML '98)*, Morgan Kaufmann, San Fransisco, CA, 82-90 (1998).
[6] Breault, J. L., Goodall, C. R., and Fos, P. J., Data mining a diabetic data warehouse, *Artifi cial Intelligence in Medicine*, **26**, 37-54 (2002).

[7] Breiman, L., Friedman, J. H., Olshen, R. A. and Stone, C. J., *Classification and Regression Trees*, Wadsworth, Belmont, (1984).
[8] Burges, C. J. C., A tutorial on Support Vector Machines for Pattern Recognition, *Data Mining and Knowledge Discovery*, **2**, 121-167 (1998).
[9] Chen, J., Li, Z. and Bian, B., Application of Data Mining in Multi-Geological-Factor Analysis, *Z. Cai et al. (Eds.): ISICA 2010, LNCS 6382*, Springer-Verlag, Berlin, Heidelberg, 402-411 (2010).
[10] Cruz, J. A. and Wishart, D. S., Application of machine learning in cancer prediction and prognosis. Review, *Cancer Informatics*, **2** 59-78 (2006).
[11] Deighton, M. and Petrou, M., Data mining for large scale 3D seismic data analysis, *Machine Vision and Applications*, **20**, 11-22 (2009).
[12] Diersen, S., Lee, E., Spears, D., Chen, P. and Wang, L., Classification of Seismic Windows Using Artificial Neural Networks, *Procedia Computer Science 00 (2011)*, 1-10 (2011).
[13] Dreiseitl, S. and Ohno-Machado, L., Logistic regression and artificial neural network classification models: a methodology review, *Journal of Biomedical Informatics*, **35**, 352-359 (2002).
[14] Eftekhar, B., Mohammad, K., Ardebili, H.E., Ghodsi, M. and Ketabchi, E., Comparison of artificial neural network and logistic regression models for prediction of mortality in head trauma based on initial clinical data, *BMC Medical Informatics and Decision Making*, **5**, 1-8 (2005).
[15] Enescu, N., Seismic Data Processing Using Nonlinear Prediction and Neural Networks, In: *Proceedings of the IEEE NORSIG Symposium*, Espoo, Finland, 1-4 (1996).
[16] Friedman, N., Geiger, D. and Goldszmidt, M., Bayesian network classifiers, *Machine Learning*, **29**, 131-163 (1997).
[17] Fung, G. and Mangasarian, O. L., A feature selection newton method for support vector machine classification, *Comput. Optim. Appl. J.*, **28**, 185-202 (2004).
[18] Guyon, I., Weston, J., Barnhill, S. and Vapnik, V., Gene selection for cancer classification using support vector machines, *Machine Learning*, **46**, 389-422 (2002).
[19] Hanley, J. A. and McNeil, B. J., The meaning and use of the area under a receiver operating characteristic (ROC) curve, *Radiology*, **143**, 29-36 (1982).
[20] Hastie, T., Tibshirani, R. and Friedman, J., *The Elements of Statistical Learning: Data Mining, Inference, and Prediction*, 2nd edn, Springer, New York (2008).
[21] Hornik, K., Stinchcombe, M. and White, H., Multilayer feedforward networks are universal approximators, *Neural Networks*, **2**, 359-366 (1989).
[22] Karatzoglou, A., Meyer, D. and Hornik, K., Support Vector Machines in **R**, *Journal of Statistical Software*, **15**, 1-28 (2006).
[23] Karayiannis N. B. and Weigun, G. M., Growing radial basis neural networks: merging supervised and unsupervised learning with network growth techniques, *IEEE Trans Neural Netw*, **8**, 1492-1505 (1997).
[24] Kass, G., 'An exploratory technique for investigating large quantities of categorical data', *Applied Statistics*, **29**, 119-127 (1980).
[25] Kerh, T., Chan, Y. and Gunaratnam, D., Treatment and assessment of nonlinear seismic data by a genetic algorithm based neural network model, *International Journal of Nonlinear Sciences and Numerical Simulation*, **10**, 45-56 (2009).
[26] Kerh T., Huang, C. and Gunaratnam D., Neural Network Approach for Analyzing Seismic Data to Identify Potentially Hazardous Bridges, *Mathematical Problems in Engineering*, **2011**, 15 pages (2011).
[27] Koukouvinos, C. and Parpoula, C., Development of a model for trauma outcome prediction: A real-data comparison of Artificial Neural Networks, logistic regression and data mining techniques, *International Journal of Biomedical Engineering and Technology*, **10**, 84-99 (2012).
[28] Koukouvinos, C., Mylona, K. and Parpoula, C., A combination of variable selection and data mining techniques for high-dimensional statistical modelling, *International Journal of Information and Decision Sciences*, **5**, 154-168 (2013).
[29] Kuehn, N., Riggelsen, C. and Scherbaum, F., Facilitating Probabilistic Seismic Hazard Analysis Using Bayesian Networks, *Seventh Annual Workshop on Bayes Applications (in conjunction with UAI/COLT/ICML 2009)*, 1-7 (2009).
[30] Leon, F. and Atanasiu, G. M., Data Mining Methods for GIS Analysis of Seismic Vulnerability, *Proceedings of the First International Conference on Software and Data Technologies (ICSOFT 2006)*, INSTICC Press, Portugal, **2**, 153-156 (2006).
[31] Loh, W. Y., and Shih. Y. S., Split selection methods for classification trees, *Statistica Sinica*, **7**, 815-840 (1997).
[32] Marketos, G., Theodoridis, Y. and Kalogeras, I.S., Seismological Data Warehousing and Mining: A survey, *International Journal of Data Warehousing & Mining*, **4**, 1-16 (2008).
[33] Masters, T., *Advanced algorithms for neural networks, A C++ sourcebook*, John Wiley and Sons, New York (1995).
[34] Mohsin, S. and Azam, F., Computational seismic algorithmic comparison for earthquake prediction, *International Journal of Geology*, **5**, 53-59 (2011).
[35] Mukherjee, S., Tamayo, P., Slonim, D., Verri, A., Golub, T., Mesirov, J. and Poggio, T., Support vector machine classification of microarray data. Technical Report AI Memo 1677, MIT, CBCL Paper No. 182 (1999).
[36] Myers, R. H., Montgomery, D. C. and Vining, G. G., *Generalized Linear Models: With Applications Engineering and the Sciences*, John Wiley and Sons, New York (2002).
[37] Pardalos, P. M., Boginski, V. L. and Vazacopoulos, A. *Data Mining in Biomedicine*, Springer, New York (2007).
[38] Pardalos, P. M., Tomaino, V. and Xanthopoulos, P., Optimization and data mining in medicine, *TOP*, **17**, 215-236 (2009).
[39] Pepe, M. S., Receiver operating characteristic methodology, *J. Am. Statist. Assoc.*, **95**, 308-11 (2000).
[40] Preethi, G. and Santhi, B., Study on Techniques of Earthquake Prediction, *International Journal of Computer Applications*, **29**, 55-58 (2011).

[41] Quinlan, J. *C4.5: Programs for Machine Learning*, Morgan Kaufmann Publisher, San Mateo, California (1993).
[42] Scherbaum, F., Delavaud, E. and Riggelsen, C.,Model Selection in Seismic Hazard Analysis: An Information-Theoretic Perspective, *Bulletin of the Seismological Society of America*, **99**, 3234-3247 (2009).
[43] Schmidt, M., Fung, G. and Rosales, R., Fast Optimization Methods for L1 Regularization: A Comparative Study and Two New Approaches, *European Conference on Machine Learning (ECML 2007)*, 1-12 (2007).
[44] Soucek, B. and the IRIS Group, *Neural and intelligent systems integration*, John Wiley and Sons, New York (1991).
[45] Tibshirani, R., Regression shrinkage and selection via the lasso, *J. Roy. Statist. Soc. Ser. B*, **58**, 267-288 (1996).
[46] Vapnik, V., *Statistical Learning Theory*, Wiley, New York, (1998).
[47] Wan, S., Lei, T. C. and Chou, T. Y., A novel data mining technique of analysis and classification for landslide problems, *Nat Hazards*, **52**, 211-230 (2010).
[48] Wang, L., Zhu, J. and Zou, H., The doubly regularized support vector machine, *Statistica Sinica*, **16**, 589-615 (2006).
[49] Witten, I.H. and Frank, E., *Data Mining: Practical Machine learning Tools and Techniques with Java Implementations*, 2nd edn, Morgan Kaufmann Publishers, San Francisco (2005).
[50] Wolpert, D. and Macready, W., *No Free Lunch Theorems for Search*, Santa Fe Institute, Technical report no. SFI-TR-95-02-010 (1995).
[51] Ye, G. -B., Chen, Y. and Xie, X., Efficient variable selection in support vector machines via the alternating direction method of multipliers, In *Proceedings of the* 14^{th} *International Conference on Artificial Intelligence and Statistics (AISTATS) 2011*, Fort Lauderdale, FL, USA, JMLR W&CP, **15**, 832-840 (2011).
[52] Zhou, X. H., Obuchowski, N. A. and Obuchowski, D. M., *Statistical Methods in Diagnostic Medicine*, John Wiley and Sons, New York (2002).
[53] Zhu, J., Rosset, S., Hastie, T. and Tibshirani, R., 1-norm support vector machines, In *Advances in Neural Information Processing Systems 16, Proceedings of the 2003 Conference* (2003).
[54] Zou, H., and Hastie, T., Regularization and variable selection via the elastic net, *J. Roy. Statist. Soc. Ser. B*, **67**, 301-320 (2005).