Dimensionality Reduction of Accident Databases for Minimal Tradeoff in Prediction Accuracy

Tatiana Tambouratzis Miltiadis Chalikias Dora Souliou and Andreas Gregoriades

Abstract—**Predicting the location and severity of accidents is of paramount importance to traffic monitoring. To this end, extensive accident data collected at a multitude of locations within a given area are analyzed. For compactness of future collection and ease of analysis, the resulting databases are investigated in search of the minimal information that is necessary for the accurate prediction of accident occurrence and severity. In this piece of research, the 2005 accident dataset collected by the Republic of Cyprus Police is employed for determining the parameters that maximally affect accident severity in the area. Dimensionality reduction is performed via traditional as well as soft computing parameter selection/ extraction methodologies. Following evaluation and comparison of the results in terms of compactness and prediction accuracy, the genetic algorithm-based methodology is found to optimally combine parameter reduction and accident severity prediction accuracy as well as portability to different prediction techniques (namely decision trees and probabilistic neural networks).**

I. Introduction

THE accurate prediction of accident location and severity is a special concern of the police and traffic authorities. To this end, extensive accident data are collected over the entire road network by police officers and traffic wardens. Unfortunately, the resulting databases are usually unnecessarily large in terms of collected parameters: some parameters may prove hardly relevant to the prediction task, while others may be redundant (e.g. measure similar accident characteristics). By removing such parameters, the dimensionality – and, thus, computational complexity - of the prediction task can be reduced with a minimal effect on prediction accuracy; in the ideal case, the prediction task is facilitated as parameters that render the prediction model unnecessarily complex are removed and the relationship between collected parameters and accident severity becomes explicit.

In this piece of research, database dimensionality reduction is performed via traditional statistical and soft computing parameter selection methodologies (generalized linear model GLM [1-2] and genetic algorithm GA [3-4], respectively) as well as via a traditional mathematical parameter extraction methodology (principal component analysis PCA [5]).

In order to accurately evaluate the effect of dimensionality reduction on accident prediction as well as its portability to different prediction/estimation techniques, five-fold cross-validation (CV) [6-7] is applied as follows:

- For the GLM, both decision trees (DT's [8-11]) and probabilistic neural networks (PNN's [12-13]) are implemented (one DT and one PNN per fold).
- For GA and PCA, DT's are utilized for the dimensionality reduction (training) stage, while PNN's are implemented during the estimation (testing) stage, again using one DT and one PNN per fold.

The results of the five folds are averaged and subsequently compared in terms of compactness and prediction accuracy, thus uncovering the strengths and weaknesses of the various pattern selection/extraction methodologies as well as of the two implemented estimation techniques.

This paper is organized as follows: section II introduces the accident database used for investigating dimensionality reduction, while section III outlines the methodologies employed for parameter selection/extraction and the techniques used to implement and evaluate accident severity prediction; section IV describes and compares the results of the various methodologies and techniques in terms of accuracy and compactness, and - finally - section V concludes the paper.

II. The Accident Dataset

The 2005 accident dataset collected by the Republic of Cyprus Police (RCP, part of the Ministry of Justice and Public Order of the Republic of Cyprus) is utilized for extracting the minimal information that is necessary for accurately predicting accident severity.

Manuscript received May 7, 2010.

T. Tambouratzis is with the Department of Industrial Management & Technology, University of Piraeus, 107 Deligiorgi St, Piraeus 185 34, Greece, and the Department of Nuclear Engineering, Chalmers University of Technology, SE-412 Göteborg, Sweden (phone: 0030-210-4142423; fax: 0030-210-4142392; e-mail: tatianatambouratzis@gmail.com).

M. Chalikias is with the Department of Business Administration, Technological Educational Institute of Piraeus, Greece, and the Department of Industrial Management & Technology, University of Piraeus, 107 Deligiorgi St, Piraeus 185 34, Greece (e-mail: mchalik@teipir.gr).

D. Souliou is with the School of Electrical and Computer Engineering, National Technical University of Athens, 9 Iroon Polytechniou St, Athens 15780, Greece, and the Department of Industrial Management & Technology, University of Piraeus, 107 Deligiorgi St, Piraeus 185 34, Greece (e-mail: dsouliou@mail.ntua.gr).

A. Gregoriades is with the Department of Computer Science & Engineering, European University Cyprus, Cyprus and the School of Management, University of Surrey, UK (e-mail: a.gregoriades@euc.ac.cy).

The original RCP dataset numbers 1407 records. Each record comprises 44 parameters, namely the 38 categorical and six continuous parameters of Table I. The first 37 categorical of Table I(a) and all of the continuous parameters of Table I(b) express the independent variables (to be used as inputs in the prediction task), while the 38th categorical parameter of Table I(a) expresses the dependent (output) variable describing accident severity. The collected values 1, 2 and 3 of the dependent variable correspond to fatal, serious and light accidents, respectively, as evaluated by the police officer(s) and/or traffic warden(s) at the site of the accident.

TABLE I
CATEGORICAL (a) AND CONTINUOUS (b) PARAMETERS OF THE RCP DATABASE

Description	Values
area code	1:2
police district	1:2
month	1:12
day	1:7
time	1:8
hit and run	1: 2
main cause of accident	1:5
residential area	1:2
second cause of accident	0:5
third cause of accident	0:5
traffic control	1:3
diagram code	1:11
fourth cause of accident	0:5
conjunction type	1:8
routes permitted	1: 3
barrier type	1:5
break lane	1:4
road works	1:2
bus stop	1:2
pedestrian crossing	1:5
light conditions	1:5
accident location	1:3
road description	1:3
pavement status	1:2
weather conditions	1:2
type of event for first accident	1:4
vehicle sequence	1:5
age	1:4
gender	1:2
driver license type	1:3
manufacture year	1:4
CC	1:10
vehicle type	1:5
type of event for second accident	0:1
vehicle license	1:2
vehicle worthiness certificate	1:2
action before accident	1:2
accident severity	1:3

(a)

Description	Values
number of vehicles involved	[1 6]
number of injured	[1 7]
road width	[0 44]
pavement width	[0 7]
speed limit	[10 100]
time for ambulance	[0 55]

Of the 1327 unique records of the RCP dataset, 786, 489 and 52 records correspond to light, serious and fatal accidents, respectively, i.e. the database is imbalanced with decreasing numbers of collected records for increasing accident severity levels.

III. PARAMETER REDUCTION

A. *Parameter Selection/Extraction Methodologies*

The backward ordinal logistic GLM is the statistical parameter selection methodology used here. The GLM is capable of selecting the minimal subset of input parameters that are highly correlated with the dependent variable but not correlated with each other. Such a selection of parameters supports the inclusion of the pertinent - but not the redundant - parameters to the prediction task.

In the last decade, the GA soft computing paradigm has become a widely used [14-15] parameter selection methodology. GA employs a population of chromosomes, with the genes of each chromosome collectively representing a potential solution to the task under optimization. The fitness value of a chromosome expresses its overall fitness in constituting a solution to the task. The repeated application of the three genetic operators, namely

- crossover between chromosomes, e.g. random or directed, single- or multiple-point swapping of homologous parts of pairs of chromosomes,
- mutation of each chromosome, e.g. random or selective modification of genes, and
- selection, e.g. roulette-wheel-based or elitistic inclusion of chromosomes into the new population,

to the chromosomes of a randomly initialized population promotes the inclusion of chromosomes of eventually increasing fitness in the evolving population. Following a sufficient number of repetitions of the aforementioned procedure (generations), and provided that the GA parameters have been assigned appropriate values, at least one chromosome of the final population constitutes a (near-) optimal solution to the task at hand. GA has the added advantage of being fairly robust to small variations in the exact values (e.g. mutation probability) and options (e.g. random or directed crossover) selected for the genetic operators.

Finally, the mathematically derived PCA methodology projects the original data onto a new set of orthogonal axes in such a manner that the original multidimensional dataset with possibly correlated parameters is linearly transformed into a novel dataset of identical dimensions but with totally uncorrelated parameters. Owing to the fact that each new axis is selected so as to maximally expose the (remaining) variability of the dataset, it is not unusual for the first few axes of the PCA mapping to account for most of its variability. As a result, a small number of PCA axes are

generally sufficient in representing the original data with minimal (and usually negligible) loss of information. It is advantageous that, as a rule, PCA-derived dimensionality reduction achieves comparable classification accuracy levels to those obtained with the original dataset, without the need to take into account the outputs of the classification task while constructing the PCA axes. A disadvantage of PCA is that the extracted parameters cannot always be translated into (combinations of) original parameters, whereby pattern selection cannot be implemented accordingly.

B. *Training and Testing the Prediction Methodologies*

The five-fold CV procedure implemented for training and testing the GLM, GA and PCA methodologies operates as follows. Initially, the dataset is partitioned into five sets or equivalent cardinality, with four sets used when training each methodology and the remaining set reserved when testing it; this process is repeated five times so that each set appears a single time as a test set and the remaining four times in the training sets. The results of five-fold CV over the five test sets are averaged and overall prediction accuracy is reported.

C. *Accident Severity Prediction Techniques*

MatLab 2007a and the Neural Network Toolbox [16] are used for implementing the two evaluation techniques, namely DT and PNN.

The DT constitutes a tree-like classification technique that recursively generates data partitions in such a manner that the separation between classes is maximized at each node of the tree. The creation of maximally separable partitions is accomplished via information gain (selection of the attribute effectuating the greatest entropy reduction in the dataset, i.e. whose value(s) best partition(s) the dataset) or some other similar criterion. A single node appears at the top level of the DT, with as many second-level nodes created as there are data partitions according to the selected criterion. The branching process continues recursively in the same manner for each node of the second level and is terminated either once each data partition corresponds to exactly one class or when no further partitioning of the data is possible.

DT learning is computationally intensive and may result in extensive trees, especially for non-binary categorical and for continuous parameters. Furthermore, and owing to the level-by-level creation of the DT, classification is not necessarily optimal [10]. In order to ensure accurate learning and satisfactory as well as efficient classification of novel patterns, DT learning is customarily either combined with or followed by DT pruning.

During classification of a novel pattern, the DT is employed as a hierarchy of easy-to-understand rules, with one rule corresponding to each DT node: beginning from the top node, the novel pattern follows the appropriate nodes and branches until it reaches a terminal node, whereby it is assigned to the class represented by that node.

The four-layered PNN non-parametrically creates non-linear decision surfaces between classes that approach the Bayes optimal for datasets of increasing cardinality. PNN training is completed after a single presentation of the training patterns, with each pattern stored in a hidden node of the PNN. The smoothing parameter $\sigma \in [0,1]$, which is largely dependent on the cardinality and nature of the dataset, determines the exact shape and the amount of smoothness of the decision surface. The PNN features speedy training and direct incorporation/deletion of novel/unnecessary training patterns; however, its size and response time are directly affected by the cardinality of the dataset.

Although setting and testing the GLM is straightforward, whereby both the DT and the PNN techniques are used for evaluating prediction accuracy of the GLM-generated reduced parameter set, the GA and the PCA methodologies require a means of evaluating the quality of prediction for different candidate solutions during training, namely the collections of parameters for the GA and the number of first components for the PCA. To this end,

- DT's are employed on the training sets of the five folds for setting and fine-tuning the two methodologies, namely selecting/extracting the parameters to be used for accident severity prediction. DT's constitute advantageous tools for the discovery of good parameter subsets as they are fast to construct, do not require explicit parameter tuning, are transparent during operation and – owing to MatLab's inbuilt pruning option – automatically reach high levels of classification accuracy.
- PNN's are used for testing the two methodologies. Although, in general, slightly less accurate than DT's, PNN's offer higher discrimination to solutions (parameter subsets) that are judged as equally accurate by the DT's and can, thus, enforce more stringent prediction accuracy requirements. The reduced (in dimensionality) training sets are used for creating and optimizing the PNN's; the most advantageous value of the smoothing parameter is determined independently for each fold. Subsequently, prediction accuracy is measured by inputting the reduced test sets of the five folds to the corresponding trained PNN's.

IV. Accident Dataset Dimensionality Reduction

Dimensionality reduction is investigated for each of the aforementioned three parameter selection/extraction methodologies. The yardstick for evaluating them constitutes the prediction accuracy obtained by the DT and the PNN techniques on the original dataset, i.e. when all the 43 input

parameters of Table I are utilized. The prediction results (confusion matrix (a) and proportion of correct, underestimated and overestimated predictions (b)) are shown in Tables II and III, respectively, demonstrating the superiority (by 1.5% overall, or 3.5, 0.9 and 1.5% for fatal, serious and light accidents, respectively) of the DT over the PNN methodology in accident severity prediction.

TABLE II
CONFUSION MATRIX (a) AND PREDICTION ACCURACY (%) (b) OF THE DT METHODOLOGY FOR THE ORIGINAL 43 INPUT PARAMETERS OF THE RCP DATABASE

Actual / Predicted	Fatal	Serious	Light
Fatal	88.0952	1.2077	0.2813
Serious	9.5238	92.9952	2.2504
Light	2.3810	5.7971	97.4684

(a)

Underestimations	Correct Predictions	Overestimations
2.4850	95.5441	1.9709

(b)

TABLE III
CONFUSION MATRIX (a) AND PREDICTION ACCURACY (%) (b) OF THE PNN METHODOLOGY FOR THE ORIGINAL 43 INPUT PARAMETERS OF THE RCP DATABASE

Actual / Predicted	Fatal	Serious	Light
Fatal	84.6154	1.2270	1.1450
Serious	5.7692	92.0245	2.9262
Light	9.6154	6.7485	95.9288

(a)

Underestimations	Correct Predictions	Overestimations
3.0897	94.0467	2.8636

(b)

A. GLM Prediction Accuracy

During five-fold CV, the training sets of the five folds are used for parameter extraction. The GLM methodology establishes 15 parameters (namely area code, day, hit and run, residential area, second cause of accident, third cause of accident, traffic control, diagram code, conjunction type, barrier type, light conditions, type of event for first accident, number of vehicles involved, number of injured and speed limit) that are statistically significant at a level of significance below 0.01, another 11 (namely routes permitted, break lane, road works, pavement status, weather conditions, vehicle type, vehicle license, vehicle worthiness certificate, action before accident, pavement width and time for ambulance) that are statistically significant at a level of significance below 0.1 but above 0.01, and, finally, three parameters (namely time, type of event for second accident and pedestrian crossing) that are marginally statistically significant at a level of significance below 0.25 but above 0.1.

Subsequently, these groups of parameters are employed for investigating the parameter selection process as follows. The first group of the 15 mostly statistically significant parameters is used in its entirety (as no difference in significance can be discerned for such low values of significance) for creating the training and test sets of the five folds and extracting prediction accuracy, evaluated by the DT and PNN methodologies. The same procedure is repeated for the first two groups of parameters together, in other words prediction accuracy is investigated on 26 parameters, as - again - no conclusive comparison of significance can be reliably performed at levels of significance below 0.1. Finally, the three remaining and marginally significant parameters are added to the 26 parameters one by one and prediction accuracy is investigated in the same manner for each addition.

TABLE IV
CONFUSION MATRIX (a) AND PREDICTION ACCURACY (%) (b) OF THE DT METHODOLOGY FOR THE 26 STATISTICALLY SELECTED PARAMETERS OF THE RCP DATABASE

Actual / Predicted	Fatal	Serious	Light
Fatal	87.5	0.7229	0.4144
Serious	7.5	96.6265	1.9337
Light	5	2.6506	97.6519

(a)

Underestimations	Correct Predictions	Overestimations
1.3571	96.9466	1.6964

(b)

TABLE V
CONFUSION MATRIX (a) AND PREDICTION ACCURACY (%) (b) OF THE PNN METHODOLOGY FOR THE 26 STATISTICALLY SELECTED PARAMETERS OF THE RCP DATABASE

Actual / Predicted	Fatal	Serious	Light
Fatal	69.0476	1.0870	1.0870
Serious	16.6667	86.1413	8.4317
Light	14.2857	12.7717	90.5565

(a)

Underestimations	Correct Predictions	Overestimations
5.9821	88.0359	5.9821

(b)

When using the training sets, highest accident prediction accuracy is observed – for both techniques - when the 26 parameters demonstrating levels of significance below 0.1 are used. The confusion matrices (a) as well as the overall proportion of correct, underestimated and overestimated predictions (b) are shown in Tables IV-V for the DT and PNN methodologies, respectively, averaged over the test sets of the five folds. The superiority (by 8.9% overall, or 17.4, 10.5 and 7.1% for fatal, serious and light accidents,

respectively) of the DT over the PNN methodology becomes even more accentuated for the reduced datasets than when no parameter reduction is performed. When compared to Tables II-III, it is observed that the selection of 26 out of the original 43 parameters, which effectuates an almost 40% reduction in the number of parameters used, produces a 1.4%, rise in prediction accuracy for the DT (0.5, 3.6 and 0.2% for fatal, serious and light accidents, respectively), which can be attributed to the improved ability of the DT to predict accident severity once irrelevant parameters are removed. The PNN, however, demonstrates a contradictory performance with a drop of 6% overall and an acute decline in the prediction of fatal, serious and light accidents (by 15.6, 5.9 and 5.3%, respectively).

B. GA Prediction Accuracy

The GA is employed for determining a (near-)minimal as well as (near-)optimal subset of input parameters that accomplish high accident severity prediction accuracy.

Each chromosome represents a possible combination of parameters, with the ith binary gene (i=1, 2, … 43) of a chromosome taking on a value of 1 (or 0) if the corresponding parameter is (not) included in the combination of parameters expressed by the chromosome. The fitness value of a chromosome expresses the power of the corresponding candidate solution to accurately as well as efficiently predict accident severity. Overall fitness is formulated as

$$fitness\,value = 0.85 \times PRA/100 + 0.15 \times (1 - PAN/43) \quad (1)$$

where PRA and PAN denote the prediction accuracy (%) and cardinality of the parameter (sub)set represented by the gene, respectively. According to (1), the smaller the combination of parameters and the larger the classification accuracy, the higher the fitness value of the chromosome; the values of 0.85 and 0.15 in (1) have been selected from the initial ranges of [0.75 0.95] and [0.05 0.25], respectively, after trial-and-error.

The following GA process is repeated 50 times for each fold. The original population comprises 30 randomly created chromosomes, whose fitness is evaluated via a DT constructed using the training set of the corresponding fold. Random single-point crossover is applied to 10 pairs of chromosomes picked according to roulette-wheel selection; subsequently, each gene of the 50 chromosomes is independently submitted to random mutation with a mutation probability of 0.05; finally, after the fitness of the 50 mutated chromosomes is re-evaluated, roulette-wheel selection is applied for determining the chromosomes to appear in the population of the next generation. This procedure is repeated for 30 generations and the chromosome of maximum fitness is selected as the optimal GA solution.

Although convergence upon a unique parameter subset is not accomplished for the specific GA set-up, the parameter subsets corresponding to the optimal chromosomes of each run consistently range in size between 18 and 25 and accomplish DT overall prediction accuracy above 95%. Tables VI-VII show accident prediction accuracy evaluated over the test sets of the five folds by the DT and the PNN techniques; a typical (in terms of both size and selected parameters) optimal chromosome has been used, with the 23 selected parameters being month, day, time, main cause of accident, residential area, conjunction type, routes permitted, barrier type, road works, bus stop, pedestrian crossing, light conditions, accident location, weather conditions, age, gender, vehicle type, type of event for second accident, vehicle worthiness certificate, number of vehicles involved, road width, speed limit and time for ambulance. For the PNN's, the optimal values of the smoothing parameter are determined using the reduced (with 23 parameters only) training sets of the five folds.

TABLE VI
CONFUSION MATRIX (a) AND PREDICTION ACCURACY (%) (b) OF THE DT METHODOLOGY FOR AN OPTIMAL GA-GENERATED PARAMETER SUBSET OF THE RCP DATABASE

Actual / Predicted	Fatal	Serious	Light
Fatal	87.1795	0	0.5487
Serious	7.6923	95.7627	2.7435
Light	5.1282	4.2373	96.7078

(a)

Underestimations	Correct Predictions	Overestimations
1.7825	96.0784	2.1390

(b)

TABLE VII
CONFUSION MATRIX (a) AND PREDICTION ACCURACY (%) (b) OF THE PNN METHODOLOGY FOR AN OPTIMAL GA-GENERATED PARAMETER SUBSET OF THE RCP DATABASE

Actual / Predicted	Fatal	Serious	Light
Fatal	73.6842	1.0929	0.8117
Serious	7.8947	91.5301	5.6818
Light	18.4211	7.3770	93.5065

(a)

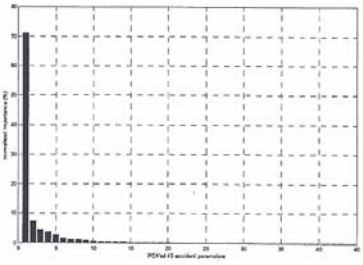
Underestimations	Correct Predictions	Overestimations
3.6275	92.0588	4.3137

(b)

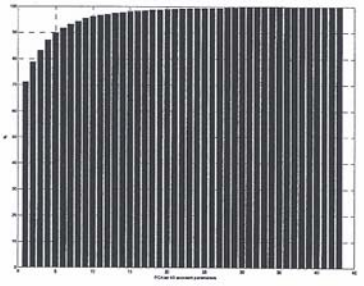
Again, prediction accuracy via the DT is found superior to the PNN by 4% overall (13.5, 4.2 and 3.2% for fatal, serious and light accidents, respectively).

When compared to Tables II-V, the removal of 20 parameters (effectuating a 46.5% reduction in parameter dimensionality) via the GA technique still allows the superior prediction of accident severity for the DT when

compared to using the entire dataset (0.5% improvement overall). The DT is less accurate than when the GLM-derived reduced dataset is used, but the difference in prediction accuracy is minor, never reaching 1% either overall or when each accident severity class is considered independently. The situation is slightly different for the PNN, which shows a 2% overall drop in prediction accuracy (9, 0.5 and 2.4% for fatal, serious and light accidents, respectively) when compared to using the entire dataset. Despite this decline, PNN performance is superior for the GA-generated than for the statistically selected reduced parameter set (improvement of 4% overall and 4.6, 5.4 and 3% for fatal, serious and light accidents, respectively).



(a)



(b)

Fig. 1 Proportion (a) and cumulative proportion (b) of total variance accounted by each component.

When examining the parameters selected by the GA, only seven, five and three appear in the group of GLM statistically significant parameters at levels of significance below 0.01, 0.1 and 0.25, respectively, in other words the GLM and the GA have only about 50% parameters in common. Further investigation is needed – in terms of GA set-up and operation – to determine to what extent the difference in selected parameters actually reflects a alternative, and perhaps more effective, parameter selection method.

C. PCA Prediction Accuracy

PCA analysis of the 43 input parameters results in 43 principal components that are orthogonal to each other and maximally expose the variability of the original dataset.

A multitude of component extraction criteria exist in the literature; these serve to determine the number of meaningful (first) components that should be retained in order for no significant loss of information of the original database to occur. Of these, the eigenvalue-one [17] criterion supports retaining the first five components, while the scree test [18], the proportion of variance and the cumulative variance [19] accounted for by the components (shown in Fig. 1) propose an extensive range of possible numbers of components, namely 1, 2, 3, 5, 7, 9 and 14.

In order to reach a consensus concerning the number of components to be retained, the DT is used for evaluating accident prediction accuracy when retaining the first one, the first two, the first three and so on until all 43 principal components; this procedure is applied to the training sets of the five folds. Following averaging, and as shown in Fig. 2, a 99.9075% prediction accuracy of is reached when using the first three components; prediction accuracy drops again until nine components are used, whereby it converges at 100% accuracy from 12 components onward[1].

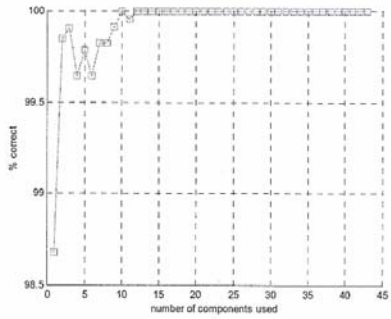


Fig. 2 Overall prediction accuracy of the DT for increasing numbers of PCA components; average over the training sets of the fie folds.

Tables VIII-IX tabulate accident prediction accuracy accomplished by the DT and PNN, respectively, when employing the first three principal components of the test sets; again, the corresponding training sets are used for setting the optimal value of the smoothing parameter. For the DT, the results are only slightly inferior to those obtained for the entire dataset in Table II (drop of 1.5% overall, and 8, 1.2% for fatal and light accidents but increase by 0.6% for

[1] The near-100% accuracy observed in Fig. 2 for increasing numbers of first PCA components is due to the evaluation of the performance accuracy using the training sets only, i.e. of the ability of the DT's to separate the accident severity classes of the training sets.

serious accidents), as expected by the low number of PCA components retained. By contrast, the PNN accomplishes equivalent accident prediction to when the entire dataset is employed, with a 4.7 and 1% rise in accuracy for fatal and serious accidents, respectively, and a drop of 0.9% for light accidents.

Efforts to duplicate these results with the parameters that are highly correlated with the first three components (namely road width, speed limit and time for ambulance for absolute correlation coefficient values greater than 0.9, residential area for absolute correlation coefficient values greater than 0.5 and police district for absolute correlation coefficient values greater than 0.4) has yielded significantly inferior accident prediction accuracy. It is clear that, in the case of this dataset, parameter selection cannot be performed based on the extracted PCA components.

TABLE VIII
CONFUSION MATRIX (a) AND PREDICTION ACCURACY (%) (b) OF THE DT METHODOLOGY FOR THE FIRST THREE PCA COMPONENTS OF THE RCP DATABASE

Actual / Predicted	Fatal	Serious	Light
Fatal	80	1.5915	0.7587
Serious	10	93.6340	3.0349
Light	10	4.7745	96.2064

(a)

Underestimations	Correct Predictions	Overestimations
2.2514	94.8405	2.9081

(b)

TABLE IX
CONFUSION MATRIX (a) AND PREDICTION ACCURACY (%) (b) OF THE PNN METHODOLOGY FOR THE FIRST THREE PCA COMPONENTS OF THE RCP DATABASE

Actual / Predicted	Fatal	Serious	Light
Fatal	89.3617	0.4082	1.0204
Serious	4.2553	93.0612	3.9541
Light	6.3830	6.5306	95.0255

(a)

Underestimations	Correct Predictions	Overestimations
2.8009	94.0954	3.1037

(b)

It is worth mentioning that the original parameters that are mostly correlated to the extracted parameters tend to be continuous rather than categorical, a phenomenon that is largely due to the ability of the continuous parameters to better distinguish between outputs (even at the cost of highly irregular separating hyperplanes), something that may not be possible for categorical parameters (especially when their values are identical for different outputs).

D. General Results

It appears overall that the DT prediction technique is better suited to pattern selection, whereby higher accident prediction accuracy is obtained when superfluous or correlated independent parameters are removed and at most 60% of the original parameters are retained. By contrast, the PNN prediction technique appears to be better tailored to parameter extraction; superior as well as significantly more efficient performance is observed when using only three PCA parameters, at the same time also accomplishing a 93% smaller PNN size, and comparable cuts in training and testing times.

Focusing upon the dimensionality reduction methodologies, the GA and PCA seem to be more robust than the GLM as far as the selection of prediction technique is concerned. As, however, parameter selection is preferable to parameter extraction for the purposes of (a) compact as well as accurate accident dataset collection, (b) ease of parameter reduction (either selection or extraction), (c) efficiency of the employed prediction techniques, and (d) understanding the characteristics of the retained parameters and their underlying relation to the dependent variable accident severity.

V. CONCLUSION

In this piece of research, the reduction of the dimensionality of the Republic of Cyprus Police accident database is performed via the statistical backward ordinal logistic - generalized linear model and the genetic algorithm-based parameter selection methodologies as well as via the principal component analysis parameter extraction methodology. Decision trees and probabilistic neural networks are employed as prediction techniques. Although principal component analysis only requires three our of the original 43 components for both perfect reconstruction of the original dataset and maximally accurate accident severity prediction, no combination of original parameters loading on these components could duplicate these results. By contrast, the genetic algorithm-based parameter selection effectuates a 46.5% reduction in parameter dimensionality of the original database, with only a 2% fall in prediction accuracy for the probabilistic neural network and a rise in prediction accuracy for the decision tree technique; furthermore, the genetic algorithm is found superior overall to the statistical backward ordinal logistic - generalized linear model.

Although a compact as well as accurate parameter subset that is capable of differentiating between accidents of varying levels of severity is put forward by the GA methodology, further analysis must be performed for determining the true effect that the (categorical/continuous) nature of the data corresponding to each parameter has on its significance for predicting accident severity and for relating

these results to those obtained from statistical parameter selection models.

REFERENCES

[1] D. W. Marquardt, "A Critique of Some Ridge Regression Methods: Comment," *Journal of the American Statistical Association*, vol. 75, pp. 87-91, 1980

[2] M. B. Priestley, *Non-Linear and Non-Stationary Time Series Analysis*, Academic Press, 1988

[3] J. H. Holland, *Adaptation in Natural and Artificial Systems*, University of Michigan Press, Ann Arbor, 1975

[4] D. E. Goldberg, *Genetic Algorithms in Search, Optimization and Machine Learning*, Kluwer Academic Publishers, Boston, MA, 1989

[5] K. Pearson, "On lines and planes of closest fit to systems of points in space," *Philosophical Magazine*, vol. 2, pp. 558-572, 1901

[6] P. A. Devijver, J. Kittler, *Pattern Recognition: a Statistical Approach*, Prentice-Hall, London, 1982

[7] G. J. McLachlan, K.A. Do, C. Ambroise, *Analyzing Microarray Gene Expression Data*, Wiley, 2004

[8] W. Y. Loh and N. Vanichsetakul, "Tree-structured classification via generalized discriminant analysis," *Journal of the American Statistical Association*, vol. 83, pp. 715-728, 1988

[9] Y.-S. Shih, "Selecting the best categorical split for classification tree," *Statistics and Probability Letters*, vol. 54, pp. 341-345, 2001

[10] L. Hyafil and R. L. Rivest, "Constructing Optimal Binary Decision Trees is NP-Complete," *Information Processing Letters*, vol. 5, pp. 15-17, 1976

[11] J. R. Quinlan, *C4.5: Programs for Machine Learning*, Morgan Kaufmann Publishers, 1993

[12] D. Specht, "Probabilistic neural networks for classification, mapping, and associative memory," in Proc. *1988 IEEE Int. Conf. on Neural Networks*, New York, U.S.A, pp. 525-532 (vol. 1), 1998

[13] D. F. Specht, "Probabilistic neural networks and the polynomial Adaline as complementary techniques to classification," *IEEE Transactions on Neural Networks*, vol. 1, pp. 111–121, 1990

[14] W. Sidlecki and J. Sklansky, "A note on genetic algorithms for large scale feature selection," pp. 88-107, in *Handbook of Pattern Recognition and Computer Vision*, C. H. Chen, L. F. Pan, P. S. P. Wang (eds.), World Scientific, 1993

[15] S. K. Pal, and P. P. Wang (eds), *"Genetic Algorithms for Pattern Recognition"*, CRC Press, 1996

[16] The MathWorks, R2007b MatLab & Simulink, Fuzzy Logic Toolbox (September 2007)

[17] H. F. Kaiser, "The application of electronic computers to factor analysis," *Education and Psychological Measurement*, vol. 20, pp. 141-151, 1960

[18] R. B. Cattell, "The scree test for the number of factors," *Multivariate Behavioral Research*, vol. 1, pp. 245-276, 1966

[19] J. O. Kim and C. W. Mueller, "*Factor Analysis, Statistical Methods and Practical Issues*", Sage, Beverly Hill, CA, U.S.A.