A Fatality Analysis Reporting System Dataset Study

Alimed Celecia Ramos
Electrical Engineering Department
PUC-Rio
Rio de Janeiro, Brazil
alimedcr22@gmail.com

Rainel Sánchez Pino
Electrical Engineering Department
PUC-Rio
Rio de Janeiro, Brazil
rainelsp89@gmail.com

Adilaraima Martínez Barrios
Electrical Engineering Department
PUC-Rio
Rio de Janeiro, Brazil
adilamtnez@gmail.com

***Abstract*— Fatality Analysis Reporting System (FARS) datasets are extensively used to analyze traffic accidents in order to improve the road safety and to design better accident prevention campaigns. This work is a case study of the FARS dataset for the year 2001, classifying the traffic accidents in function of the involved injury producing capacity. The challenges of this dataset are the mixture of categorical, real and integer features and the class disbalance, with almost four time of features for a class than the other. In order to respond to that problems then are applied 9 different classifiers: Linear Discriminant Analysis (LDA), Naïve-Bayes, Decision Trees, SVM linear, SVM with gaussian kernel, Multilayer Perceptron Neural Network (MLP), Radial Basis Function Neural Network (RBF), Probabilistic Neural Network (PNN), and Random Forest. Those models were combined with the class balancing algorithms Smote and ENN, in conjunction with a Genetic Algorithm for feature selection in a wrapper approach. The best model obtained an accuracy of 99.82%.**

Keywords*— *FARS dataset; Smote; class imbalance problem; feature selection; Genetic Algorithm

I. INTRODUCTION

Traffic accidents are one of the main death causes in the majority of the countries, being an evitable problem due to its main causes (fundamentally related with drivers or pedestrians negligence). The high number of casualties on traffic accidents (1,250,000 worldwide [1]) can be avoided by identifying the main causes of those accidents, and posteriorly prevent them through road safety campaigns. Significantly, the lesions caused by traffic accidents are the eighth cause of death worldwide, and the first among young people from 15 to 29 years. Current trends indicate that if urgent measures are not taken, by the year 2030, traffic accidents will become the fifth cause of death in the world.

Several Statistic International and National Organizations offer the annual data about traffic accidents. Those data provide a great opportunity to apply Machine Learning and Data Mining algorithms to answer some of the questions that could appear in the process of understanding the accidents: which their main causes are, and if it's possible to predict the outcome of the accident in the sense of health consequences for the persons involved.

The last question can be answered using the Fatality Analysis Reporting System (FARS) that annually is shared by the National Highway Traffic Safety Administration of the United States. These data cover characteristics of traffic accidents (quantity of passengers in the vehicle, alcohol or drugs involvement, fatalities, injuries, among others) since 1975, and can be freely accessed by anyone through the internet.

The dataset FARS of the year 2001 have been already employed in several studies for testing Machine Learning and Data Mining algorithms [2], [3], [4], [5], [6]. It is composed by attributes describing characteristics of traffic accidents and its outcomes focused on injury levels. State of the art works, indicated in the references ([3], [4], [5], [6]), employ a feature generation and data filtering named l-diversity combined with anatomization and different classifiers (specifically Support Vector Machines (SVM) and k-Nearest Neighbor (kNN)).

In this work the objective is to classify traffic accidents into injury or not injury producers. This two classes are a result of a new distribution of several classes of the dataset, resulting in a disbalanced representation, with much more examples for injury producers that for the non-injury producers. In this case, were applied two methods to tackle the balance problem. Furthermore, it is applied also a feature selection algorithm for the optimization of the data representation. Also, were tested several classification algorithms, including Linear Discriminant Analysis (LDA), Naïve-Bayes, Decision Trees, SVM linear, SVM with gaussian kernel, Multilayer Perceptron Neural Network (MLP), Radial Basis Function Neural Network (RBF), Probabilistic Neural Network (PNN), and Random Forest.

The remainder of this paper is organized as follows: next section describes the proposed methods for the dataset preprocessing and classification, including also a detailed description of the dataset. Section III presents the dataset employed and the transformation preliminary manipulations to classes reorganization. The results obtained for the different applied algorithms and methods are described in Section IV. Finally, Section V provides the conclusions of our work and possible future improvements.

II. METHODS

As stated in the Introduction, several classification algorithms were applied for the classification task. Also, some

methods for feature selection and class balance were tested to improve the performance of the system. All the employed algorithms are briefly introduced in this section.

A. Naïve-Bayes

The Naive Bayes algorithm is an intuitive method that employs the probabilities of attributes belonging to each class to make a prediction based on Bayes Theorem. In its design, it simplifies the calculation of probabilities by assuming that the probability of each attribute belonging to a given class value is independent of all other attributes. The algorithm then predicts a class, given the a priory probabilities of the set of features. This is a strong assumption but results in a fast and effective method [7].

B. LDA

Linear Discriminant Analysis (LDA) [8] is based on the assumption that each probability density functions of the class can be addressed as a normal density and have the same covariance. The objective of LDA is to determine a hyperplane that separates the data into their classes. This hyperplane is obtained by the search of the projection that maximizes the distance between the mean vector of each class and minimize the interclass variance. A new data is classified determining the highest probability density function.

C. SVM linear and gaussian kernel

A Support Vector Machine (SVM) [9] is a classifier that employs a separating hyperplane that maximizes the margins between the data classes mapped into a space of higher dimension by the "kernel trick". This maximization leads to an increase of the generalization capacity of the model, and the use of different kernels helps to create nonlinear decision boundaries. In addition, the SVM is known by its immunity to overfitting and the "curse of dimensionality". In this work the parameter of the SVMs are the C.

D. Decision Trees

Decision tree algorithm [10] is a data mining induction technique that recursively partitions a data set of records using depth-first greedy approach until all the data items belong to a particular class. A decision tree structure is made of root, internal and leaf nodes, and is used as a track way in classifying unknown data records. At each internal node of the tree, a decision of best split is made using impurity measure.

E. Random Forest

Random Forest [11] is a tree-based ensemble with each tree depending on a collection of random variables. Fits many classification trees to a data set, and then combines the predictions from all the trees. The generalization error of a forest of tree classifiers depends on the strength of the individual trees in the forest and the correlation between them.

F. MLP Neural Network

The MLP networks [12] are neural networks of feed-forward topology that contain an input layer, one or more hidden layers and an output layer. The layers are constituted by neurons (processors) that have computational capacities. The number of neurons in the output layer is determined by the required dimensionality of the desired response. For the choice of the number of hidden layers and the number of neurons of the same, there are no defined rules. These two aspects determine the complexity of the network. The activation functions of neurons must be non-linear and differentiable. Nonlinearity serves to separate patterns that are not linearly separable and differentiation allows the calculation of the decreasing gradient of the function, thus directing the adjustment of neuron weights during training. The specification of the synaptic weights that interconnect the neurons in the different layers of the network involves the use of supervised training algorithms. The most used training algorithm in these networks is the backpropagation algorithm, based on learning by error correction. In their implementation was defined the activation function as linear, and the number of neurons of the hidden layer were varied in order to find the best structure.

G. Radial Basis Function Neural Network

RBF is also composed of several layers and uses supervised learning. Its main feature is the use of radial base functions in all nodes of the hidden layer, which, instead of using as a function argument the scalar product between the values of the input register and the values of the register of weights of the neuron, uses the distance between the input vectors and their center [13]. Like the MLP network, the RBF network is one of the neural networks that have been successfully applied to many classification tasks. The RBF network structure is formed by an input layer, only a hidden layer, which applies a nonlinear transformation of the input space to a high dimensional space, and the output layer, which applies a linear transformation in space by providing an output to the network. In addition, RBF networks are also feedforward and fully connected. For the design of the algorithm the varied two parameters: the maximum number of neurons and the spread of the gaussian function.

H. Probabilistic Neural Networks

PNNs are neural networks inspired by the Bayesian classifiers, those models accomplish the approximation to an optimal Bayes under the conditions of an training set big enough. This Neural Network is composed of two hidden layers: one named the pattern layer that will have a processor for every vector of entry and its activation functions are a bell shape function that scales the variable nonlinearity, and other named the summation layer that computes the sum of all the processors of a certain class. The output layer computes the decision of the belonging to any of the classes. Weights are assigned, not trained and existing weights will never be alternated, and only new vectors are inserted into weight matrices with training. The parameter that controls the algorithm is the spread of the gaussian function.

I. Class balance algorithms

1) Smote

Regarding the generation of synthetic data, the oversampling technique (Smote) [14] has shown great success in several applications. The Smote algorithm creates artificial data based on the similarity feature of existing minority class examples.

Specifically, for a subset $S_{min} \in S$ consider the nearest k for each example $x_i \in S_{min}$, for any specified integer k; the nearest k is defined as S_{min} element if the Euclidean distance between it and x_i into consideration exhibits a smaller magnitude along the n-dimensional space X. To create a synthetic example, select is randomly one of its nearest k neighbors, then multiplies the difference between the corresponding value by a random number between [0,1], and finally adds this new value to the subset S_{min}

$$x_{new} = x_i + (y_i - x_i)\delta \quad (1)$$

where $x_i \in S_{min}$ is an example of the minority class under consideration, y_i is one of its nearest k, x_i: $y_i \in S_{min}$, and $\delta \in [0,1]$ is a random number. Therefore, the result of the synthetic example generated is a point along the line joining the point x_i into consideration with one of the nearest randomly selected k.

2) ENN

The ENN (Edited Nearest Neighbors) method was created by Wilson in 1972 [15] and consists of the elimination of all objects that are misclassified by their k - nearest neighbors. This process is basically based on:

For each point $x \in$ Dataset

1- One discovers who are the nearest k-neighbors from x.
2- It is verified the classification of all k-neighbors of x.
3- If the classification of x is different from the classification of all its k-neighbors, this point is eliminated.

The application of the ENN method is done after the application of the Smote method. ENN is used to remove examples from both classes. So any example whose classification is different from its three closest neighbors is eliminated.

J. Genetic Algorithm for feature selection

Genetic Algorithm (GA) is an optimization method inspired by natural evolution and genetic recombination. This technique is built on the principle defined by Charles Darwin of reproduction and survival of the fittest.

For feature selection, the chromosome is usually coded as a binary array with as many genes as features in the problem. The features with an active value in the chromosome are the ones selected for the subset validated by the learning algorithm. In our case the genetic algorithm was configured with tournament selection, scattered crossover and gaussian mutation; the population size and the number of generations were varied in the experiments as in [16].

III. DATASET

The FARS dataset describes characteristics of traffic accidents (as quantity of passengers in the vehicle, alcohol or drugs involvement, fatalities, injuries, among others) during the year 2001 in the United States roads. It was compiled and shared by the National Highway Traffic Safety Administration of the United States.

The dataset present the following characteristics:

TABLE 1. CHARACTERISTICS OF THE SELECTED DATASETS

Instance number	100,968
Number of attributes	29
Characteristic of attributes	Integer, Real, Categorical

The features describe fundamental characteristics of the traffic accidents. It contains: age of the involved person, sex, sitting position, if it was wearing the belt, alcohol or drugs involvement, race, among others.

The dataset present eight labels, which are distributed as can be seen in the next graphic:

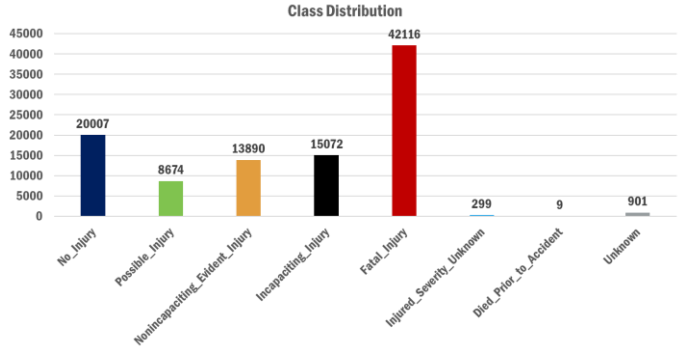

Figure 1. Class distribution

As it can be seen the classes are highly unbalanced, with a smaller representativity for the Injury Severity Unknown, Died Prior to Accident, and Unknown classes. In the state of the art literature, the form of solution of these problem is with a new labels definition. In this case the four classes Possible Injury, Nonincapaciting Evident Injury, Incapaciting Injury, and Fatal Injury are grouped under the new label Injury, and No Injury is defined as the other class. The three less represented classes are eliminated from the problem. As result, there are 71,078 instances of the Injury class and 20,007 for the No injury class.

IV. RESULTS AND DISCUSSIONS

In order to test all the algorithms and validate their usefulness for this dataset the instances were divided as follow:

TABLE 2. DATASET DIVISION

Groups		**Number of Instances**
Training Set (70%)	Training Set (60%)	38,255
	Validation Set (40%)	25,501
Test Set (30%)		27,326

Firstly, there were applied on the raw data the nine classifier algorithms. There were tested exhaustively the possible values of the correspondent parameters for each one. This was implemented in order to obtain the best possible structure for each model as a for loop testing several values in a given numerical space or limit. Each model is trained with the training set, and their results validated in the validation set. The test set is left untouched until the last test with the best obtained model. The performance metric used to evaluate and compare the models is the percentage of correctly classified patterns.

$$P = 100 * \frac{C_C}{C_T} \quad (2)$$

Where P is the performance given in percentage, C_C the number of correct classifications and C_T the total number of patterns. The algorithms were implemented using Matlab2016b and Python 3.6 distributions.The resultant accuracies are shown in the next table.

TABLE 3. PERFORMANCES OBTAINED USING THE RAW DATA

Classifiers	Performance (%)
Naïve-Bayes	92.24
LDA	92.31
Decision Tree	98.98
MLP Neural Network	77.90
RBF Neural Network	87.25
PNN	97.13
SVM linear	95.41
SVM gaussian kernel	94.05
Random Forest	99.22

In general, the results for all the classifiers were relatively good, being the outstanding ones the Decision Tree and the Random Forest, which can be a clue of the positive employment of entropy and information gain metrics for feature selection heuristics. The worst performance was offered by the MLP. This is a product of the description of the dataset, in where the categories of categorical features are offered as numbers, which is knowingly harmful for MLPs. The best practice in those cases is to conduct a binarization of the data, in which the number of bits are equal to the number of categories of each feature, and when the category is present the position is set to one. This approach certainly would provide better results for the MLP model, but in this dataset there are 24 categorical features with different number of categories (ranging from 3 to 50). As a result, the number of entries of the Neural Network will be over 200, which turns it training prohibitively costly. This, summed to the great results of the best models make the authors decide to eliminate the MLP for further tests, focusing then in more promising and faster algorithms.

It is known that one of the problems of the dataset are the disbalance of the classes (with almost four times the quantity of examples for injury that for non-injury). To solve this issue there was observed the possibility of application of several algorithms. As a result and observing the conclusions of [17], was selected the Smote oversampling method combined with the ENN undersampling method. The results of the application of the methods (first with Smote and secondly with Smote+ENN) are described in Tables 4 and 5.

TABLE 4. PERFORMANCE OBTAINED USING CLASS BALANCING METHODS

Classifiers	Smote Performance (%)	Smote+ENN Performance (%)
Naïve-Bayes	92.25	91.98
LDA	92.29	91,75
Decision Tree	99.19	99.67
RBF Neural Network	83.86	82.37
PNN	97.96	98.84
SVM linear	97.39	97.98
SVM gaussian kernel	95.47	95.86
Random Forest	99.40	99.75

The results demonstrate that are algorithms that present a greater improvement than others after the employment of those balancing methods. Significantly, PNN and SVM linear improved in several percentage points their accuracy, with a moderate improvement for the Decision Tree and Random Forest. On the opposite, Naïve-Bayes and LDA suffered a decay in their accuracies for both methods.

The addition of the ENN represented an improvement of the algorithm, given that helped to eliminate those examples in conflict with the classes of the majority of it neighbors. Except the classifiers with a worse performance, the others improved their capacity of classification of examples. The best performance is offered by the Random Forest, with a 99.75% of the examples classified correctly.

Given those results, it was selected the Random Forest model for further improvement, employing it in combination of a Genetic Algorithm in a wrapper approach (superior method found in the literature for feature selection [16]). In the case the wrapper optimizes the accuracy of the classifier selecting the best subset of features in function of which genes of the chromosome are set to one. The GA was designed with 10 generations and 20 individuals, with a crossover probability of 0.8 and a mutation probability of 0.1. Figure 2 displays the evolution results in each generation.

The evolution process improved slightly the accuracy of the model, with the best value of 99.84% of correct classifications. That is the best obtained results, so the model then is selected for process the test set that was left untouched. That last test is represented in Table 5.

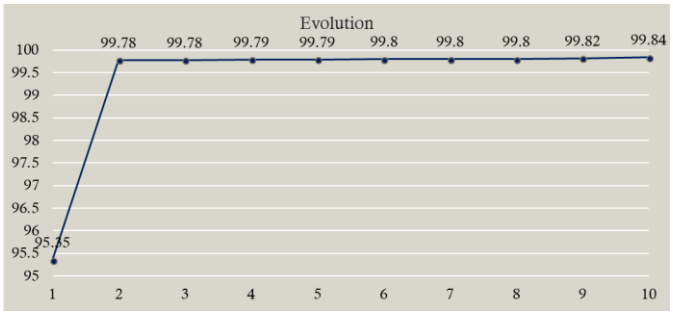

Figure 2. Evolution result

TABLE 5. RESULT OF THE PROPOSED MODEL FOR THE TEST SET

Preprocessing methods	Classifiers	Features	Accuracy (%)
Smote+ENN+GA	Random Forest	19	99.82

The accuracy of the model is very good. In order to validate the results then the proposed model can be compared to the results obtained in the state of the art literature. The results are shown in the next table:

TABLE 6. STATE OF THE ART RESULTS COMPARISON

	Model	Performance (%)
Proposed Method	Smote+ENN+GA+RF	99.82
[6]	ldiversity+anatomization+SVM linear	99.99
	ldiversity+anatomization+SVM rbf	99.98
[4]	ldiversity+anatomization+kNN, n=3	99.99
	ldiversity+anatomization+kNN, n=5	99.99
	ldiversity+anatomization+kNN, n=7	99.99
	ldiversity+anatomization+kNN, n=9	99.99

The observation of the results of the Mancuhan works shows that, even with a good performance, the model is below by almost 0.2 percentage points of his results. This can be mainly a cause of the superiority of the preprocessing technique employed by him, which selects features and generates diverse new examples, being named l-diversity and anatomization.

V. CONCLUSIONS

This work presented the case study of the 2001 FARS dataset. The objective was to find the best model for classifying the accident in injury producers or not, which can be very helpful in the analysis of this data and in the design of road safety campaigns. With this purpose were applied 9 classifications methods (LDA, Naïve-Bayes, Decision Trees, SVM linear, SVM with gaussian kernel, MLP, RBF, PNN, and Random Forest), validating exhaustively their design parameters to give the best possible model. Also, the class disbalance problem was treated using a combination of Smote and ENN methods, which improved the performance of the majority of the models. The best model then was selected for a feature selection algorithm that integrates an GA to the process. In this case it helped to select the subset of features that described the data in a best way for the Random Forest classification. The result of the best model was of 99.82 of correctly classified examples.

As future works can be implemented other class balance heuristics as Tomen Links. In addition, an ensemble of other classifiers (Random Forest is an ensemble of only decision trees) can result on a more robust model. Finally, the GA as feature selector can be applied with others classifiers, which certainly will result in an improvement of their performances.

REFERENCES

[1] WHO, ed. (2015). "WHO Report 2015: Data tables" (PDF) (official report). Geneva: World Health Organisation (WHO)

[2] "Knowledge Extraction based on Evolutionary Learning" [Online]. Available: http://sci2s.ugr.es/keel/index.php. [Accessed: 16-March-2018].

[3] Mancuhan, K. and Clifton, C. "K-Nearest Neighbor Classification Using Anatomized Data." *CoRR*, abs/1610.06048, 2016.

[4] Mancuhan, K. and Clifton, C. "Statistical Learning Theory Approach for Data Classification with ℓ-diversity.", *SDM*, 2017.

[5] Mancuhan, K., and Clifton, C. "Instance-Based Learning with l-diversity". *Transactions on Data Privacy*, vol. 10, 2017.

[6] Mancuhan, K., and Clifton, C. "Support vector classification with ℓ-diversity". *Computers & Security*, 2018.

[7] Zhang, J., Kang, D.-K., Silvescu, A., and Honavar, V. "Learning accurate and concise naïve Bayes classifiers from attribute value taxonomies and data". *Knowledge and Information Systems*, Vol. 9(2), 2006.

[8] S. Mika, G. Ratsch, J. Weston, and B. Scholkopf, "Fisher discriminant analysis with kernels," *Neural Networks Signal Process. IX, 1999. Proc. 1999 IEEE Signal Process. Soc. Work.*, pp. 41–48, 1999.

[9] B. Schölkopf and A. J. Smola, *Learning with kernels : support vector machines, regularization, optimization, and beyond*. MIT Press, 2002.

[10] J. Ross Quinlan. *C4.5: Programs for Machine Learning*. Morgan Kaufmann Publishers Inc., San Francisco, CA, USA, 1993.

[11] L. Breiman. “Random Forests”. *Machine Learning*, 2001.
[12] R. Rojas, *Neural networks : a systematic introduction*. Springer-Verlag, 1996.
[13] Claudemir Braga Carvalho, *Redes Neurais Artificiais: Teoria e Aplicações*, Livro Técnico e Científico, Rio de Janeiro, 2000.
[14] N. V. Chawla, L. O. Hall, K. W. Bowyer, and W. P. Kegelmeyer, “SMOTE: Synthetic Minority Oversampling Technique”. *Journal of Artificial Intelligence Research*, vol 16, 2002.
[15] D. Wilson, “Asymptotic Properties of Nearest Neighbor Rules Using Edited Data”. *IEEE Transactions on Systems, Man, and Cybernetics*, 1972.
[16] A. Celecia, R. González, and M. Vellasco, “Feature Selection Methods Applied to Motor Imagery Task Classification,” in *LA-CCI 2016 Latin American Conference on Computational Intelligence*, 2016.
[17] André Spinelli Schavioni, *Um estudo comparativo de métodos para balanceamento do conjunto de treinamento em aprendizado de redes artificiais*, Monografía, Lavras, 2010.