Accident Analysis and Prevention 88 (2016) 37–51

Contents lists available at ScienceDirect

Accident Analysis and Prevention

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

Bayes classifiers for imbalanced traffic accidents datasets

Randa Oqab Mujalli[a,*], Griselda López[b], Laura Garach[b]

[a] Department of Civil Engineering, The Hashemite University, 13115 Zarqa, Jordan
[b] Department of Civil Engineering, University of Granada, ETSI Caminos, Canales y Puertos, c/ Severo Ochoa, s/n, 18071 Granada, Spain

ARTICLE INFO

Article history:
Received 25 April 2015
Received in revised form 24 October 2015
Accepted 2 December 2015
Available online 20 December 2015

Keywords:
Bayesian networks
Traffic accidents
Urban area
Imbalanced data set
SMOTE

ABSTRACT

Traffic accidents data sets are usually imbalanced, where the number of instances classified under the killed or severe injuries class (minority) is much lower than those classified under the slight injuries class (majority). This, however, supposes a challenging problem for classification algorithms and may cause obtaining a model that well cover the slight injuries instances whereas the killed or severe injuries instances are misclassified frequently. Based on traffic accidents data collected on urban and suburban roads in Jordan for three years (2009–2011); three different data balancing techniques were used: undersampling which removes some instances of the majority class, oversampling which creates new instances of the minority class and a mix technique that combines both. In addition, different Bayes classifiers were compared for the different imbalanced and balanced data sets: Averaged One-Dependence Estimators, Weightily Average One-Dependence Estimators, and Bayesian networks in order to identify factors that affect the severity of an accident. The results indicated that using the balanced data sets, especially those created using oversampling techniques, with Bayesian networks improved classifying a traffic accident according to its severity and reduced the misclassification of killed and severe injuries instances. On the other hand, the following variables were found to contribute to the occurrence of a killed causality or a severe injury in a traffic accident: number of vehicles involved, accident pattern, number of directions, accident type, lighting, surface condition, and speed limit. This work, to the knowledge of the authors, is the first that aims at analyzing historical data records for traffic accidents occurring in Jordan and the first to apply balancing techniques to analyze injury severity of traffic accidents.

1. Introduction

Reducing the severity of accidents is an effective way to improve road safety (Qiu et al., 2014). Recent road traffic safety studies have focused on analysis of risk factors that affect fatality and injury level (severity) of traffic accidents. However, many risk factors are waiting to be discovered or analyzed (Kwon et al., 2015).

Traffic accidents are considered one of the most important and dangerous problems that encounters societies all around the world where it consumes many human and monetary resources. World Health Organization (WHO) statistics indicated that traffic accidents fatalities are estimated to be 1.2 million persons annually worldwide, as well as resulting in 20–50 million injuries. Cost of traffic accidents is estimated to be 518 billion US dollars representing (1–3%) of Gross Domestic Product (GDP) worldwide (WHO, 2013).

Jordan is considered a developing country which has both rapid population and vehicles growth; population statistics of 2013 issued by Department of Statistics (DOS) indicated that Jordan has 6.53 million inhabitants with 1,263,754 registered vehicles (1 vehicle/5 persons) (DOS, 2013). According to Police Traffic Department (PTD) reports for 2013; 107,864 traffic accidents occurred in Jordan with 768 fatalities, 2258 severe injuries and 13,696 slight injuries. A percentage of 94.74% of these accidents were collisions,[1] resulting in 43% of fatalities and 50% of severe injuries (PTD, 2013). Also, 69% of traffic accidents and 71% of collisions occurred in the capital city of Amman, which is considered an urban area having nearly 39% of Jordan's population (2,528,500 inhabitants). In addition, the cost of traffic accidents in Jordan, using unit cost approach to estimate traffic accidents cost in a socioeconomic perspective, is estimated to be 365 million US dollars (PTD, 2013). It is worth noting that Jordan's GDP for 2013 is estimated to be 33.641 billion US dollars, of which the cost of traffic accidents represents 1.2% (DOS, 2013).

* Corresponding author.
E-mail addresses: randao@hu.edu.jo, rmujalli@hotmail.com (R.O. Mujalli).

[1] Collisions exclude all of: run-off-road accidents, pedestrian related accidents and property damage only accidents.

http://dx.doi.org/10.1016/j.aap.2015.12.003
0001-4575/

Urban and rural accidents characteristics are different (Khorashadi et al., 2005; Theofilatos et al., 2012). Khorashadi et al. (2005) identified significant differences between urban and rural accidents due to differing driver, vehicle, environmental, road geometry and traffic characteristics. Moreover, they estimated that the severe/fatal injury is nearly eight times more likely to occur in an urban area and about 2.5 times more likely in a rural area than other types of injures (i.e. no injury, complaint of pain, or visible injury). Theofilatos et al. (2012) investigated road accident severity with particular focus on the comparison between inside and outside urban areas. They found that factors affecting road accident severity inside urban areas included young driver age, bicyclists, intersections, and collision with fixed objects, whereas factors affecting severity outside urban areas were weather conditions, head-on and side collisions. This demonstrated the particular road users and traffic situations that should be focused on for road safety interventions for the two different types of networks (inside and outside urban areas).

Many modelling techniques have been in use to analyze the injury severity of traffic accidents. The most used models were the logit and probit (Al-Ghamdi, 2002; Milton et al., 2008; Savolainen et al., 2011; Mujalli and De Oña, 2012). However, most of them have their own model assumptions and pre-defined underlying relationships between dependent and independent variables (Chang and Wang, 2006). Recently, many researchers have used methods based on data mining techniques. For example, association rules (Pande and Abdel-Aty, 2009; Montella et al., 2012) or Decision Trees (López et al., 2012a; Abellán el al., 2013; De Oña et al., 2013) have been used for identifying accident patterns. Bayesian networks (BNs) have also been used to study traffic accidents' severity. De Oña et al. (2011) employed BNs to model the relationship between injury severity and variables related to driver, vehicle, roadway, and environment characteristics. They concluded that BNs could be used for classifying traffic accidents according to their injury severity. In addition, Mujalli and De Oña (2011) presented a simplified method based on BNs and variable selection algorithms to predict the injury severity in a traffic accident. Recently, Kwon et al. (2015) used two classification methods, the Naive Bayes and the Decision Tree classifier, for the ranking of risk factors.

Traffic accidents datasets usually have fewer records for fatal and severe injury accidents than for slight injury accidents (Montella et al., 2012). A dataset is considered to be imbalanced if one of the classes (called a minority class) contains a much smaller number of examples than the remaining class (majority class) (Stefanowski and Wilk, 2008). According to Li and Sun (2012) if the proportion of minority class samples constitutes less than 35% of the dataset, the dataset is considered to be imbalanced. Data mining algorithms when learning from imbalanced data tend to produce high predictive accuracy over the majority class, but poor predictive accuracy over the minority class (Thammasiri et al., 2014). Many solutions have been proposed to this problem which can be categorized into two major groups (López et al., 2012b): the internal approaches that create new algorithms or modify existing ones, and the external approaches that preprocess the data in order to diminish the effect of the class imbalance. The pre-processing approach (or resampling techniques) seems to be the more straight forward approach that has greater promise to overcome the class imbalance problem (Thammasiri et al., 2014).

Resampling techniques can be categorized into three groups: the first group consists of the under-sampling methods, which aim to balance the class populations through removing data samples from the majority class until the classes are approximately equally represented. Under-sampling methods randomly eliminate instances from the majority class until a required degree of balance between classes is reached. The second group includes the oversampling methods, which aim to balance class populations through creating new samples from the minority class and adding them to the training set. Finally, the third group comprises the mix methods, which combine both sampling approaches, integrating oversampling of selected minority class instances with removing the most harmful (i.e. noise, and borderline instances that are close to the boundary between the positive and negative classes regions) (Stefanowski and Wilk, 2008; Błaszczyński and Stefanowski, 2015).

In this work, factors affecting injury severity of urban and suburban traffic accidents in Jordan are analyzed. For this purpose, Bayes classifiers are used in the original dataset and in the three balanced datasets (balanced with random under-sampling, with oversampling and with mix methods). Finally, the models developed are compared, and the results for the best model are described.

The paper is organized as follows: Section 2 presents the methodology, the data used, a brief description of Bayes classifiers used, and a description of the performance measures used to evaluate the models. In Section 3, the results and their discussion are presented. Finally, conclusions are given in Section 4.

2. Methodology

In this paper, an imbalanced data set was first obtained and used to develop models applying different popular Bayes classifiers: Efficient Lazy Elimination for Averaged One-Dependence Estimators (AODEsr) (Zheng and Webb, 2006), Weightily Averaged One-Dependence Estimators (WAODE) (Jiang and Zhang, 2006) and Bayesian networks (BNs), where different scores and search algorithms were employed for BNs. Moreover, three balanced datasets were created from the imbalanced data set using three balancing techniques: random under-sampling, oversampling and mix sampling. The same Bayes classifiers used to develop models form the imbalanced data set were also used to develop models from the three balanced data sets. Furthermore, Bayes classifiers were used to analyze injury severity of collisions on urban and suburban roads. The developed models were compared to each other using 10-folds cross validation method, where each data set was first divided into 10 subsets, nine were used to train the model and the remaining one subset was used to test the model. The process was repeated ten times and the average was obtained. As a result, 11 models were developed and compared. Fig. 1 shows the procedure employed.

2.1. Data

Records for traffic accidents which occurred on urban and suburban roads in Jordan were obtained from the Jordanian Police Traffic Department (PTD) for a period of 3 years (2009–2011). The total number of accidents obtained for this period was 49,693. Considering that the main objective of this study was to identify the key factors that contribute to the occurrence of a specific severity in collisions; accidents with property damage only (PDO), Pedestrian and Run-Off-Road were excluded. In this study, only accidents with collisions were analyzed, and as a result, the total number of records used was 16,815.

To identify the main factors that affect urban and suburban collisions severity, fourteen independent variables were analyzed (see Table 1). The variables chosen were based on variables available in the original dataset and the variables used in literature (Theofilatos et al., 2012; Pahukula et al., 2015). The data included variables describing the prevailing conditions at the time of the occurrence of the accident:

- Roadway information: characteristics of the roadway on which the accident occurred such as number of directions, number of lanes, horizontal alignment, grade, pavement type, and pavement surface condition.

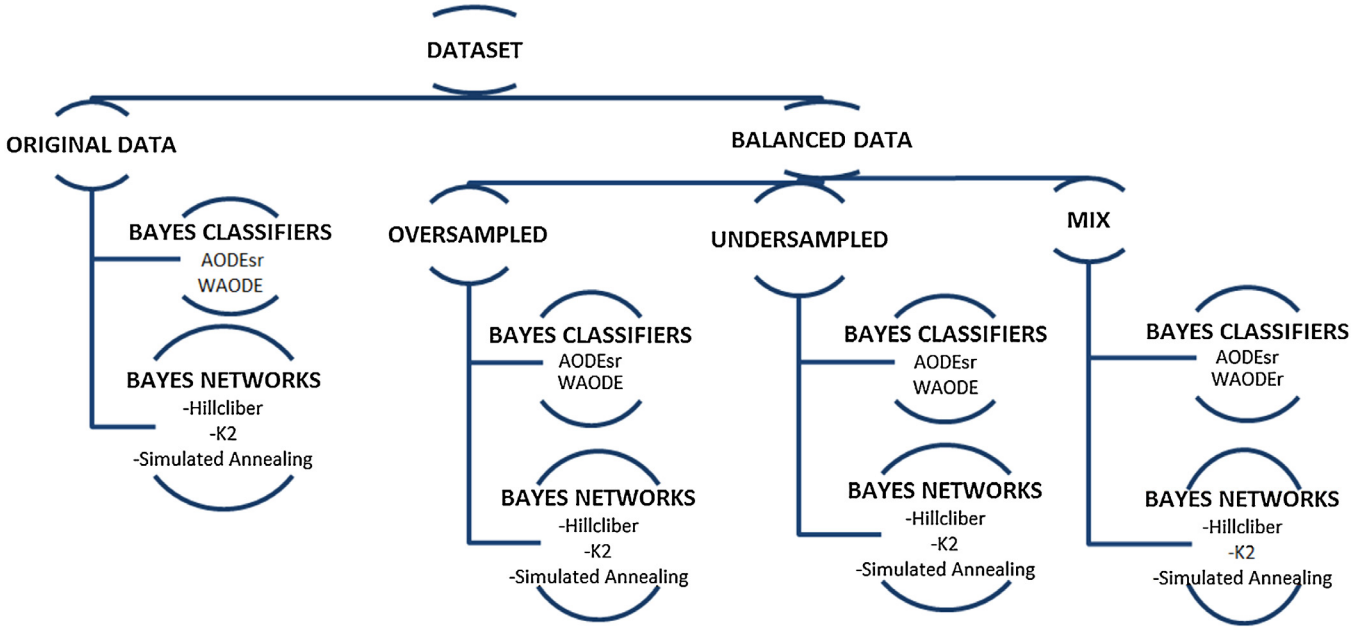

Fig. 1. The procedure employed in this study.

- Context information: weather and lighting conditions when the accident occurred.
- Accident information: contributing circumstances such as type of accident and accident pattern.
- Vehicle data: number of vehicles involved.

The class variable was the resulting severity of the accident. Following previous studies (Chang and Wang, 2006; De Oña et al., 2013; Abellán et al., 2013) the injury severity was determined according to the level of injury to the worst injured occupant. Herein, the severity was categorized in to two levels of severity: accidents with slight injuries (SLIG) and accidents with killed or severe injuries (KSEV).

The original distribution of records (also called instances) was 13,725 slight injuries and 3090 killed or severe injuries. The target variable (severity) was predominantly imbalanced, with the majority of instances belonging to SLIG (77%), and only a small percentage of KSEV (23%).

2.1.1. Data preprocessing

The variables obtained from the PTD were preprocessed prior to analysis where they were first discretized into distinct values following previous studies (Simoncic, 2004; Helai et al., 2008; De Oña et al., 2011). The unsupervised variable filter for replacing missing values was used to deal with missing data. The filter replaces a missing data with the mean if the variable was numeric or mode if the variable was nominal of all known values of that attribute in the class where the instance with missing data belongs.

Overall, out of fourteen independent variables, the following variables: PAT, TRAME, GRADE, SPE and DIR were used as they appeared in the original dataset. The rest of the values of variables were discretized in order to enable working with them. For instance, in the original dataset, the variable ACT had twelve categories, six of them have been grouped in only one category; collision with fixed object (this category included collision with guardrail, barrier, concrete barrier, pole, parked vehicle and traffic control device). Other variables such as PAV had five categories (asphalt, concrete, dirt, gravel and metal), and were grouped into three categories. Table 1 gives a description of the variables used for the analysis and their distribution between classes of severity.

2.1.2. Re-sampling techniques

A dataset is said to be imbalanced, if the number of instances in each category of the target variable is not approximately equal (Crone and Finlay, 2012). The classification problems based on imbalanced data occur often in applications when the events of interest are rare, such as the resulting outcomes of severe injuries or being killed in a traffic accident.

The class imbalance problem exists in many fields, and it was found that it caused deterioration in the performance of machine learning methods, especially classifiers performance, since they assume a balanced dataset to exist (Japkowicz, 2000). Examples of such problems were encountered in many fields such as; in flight helicopter gearbox fault monitoring (Japkowicz et al., 1995), the detection of fraudulent telephone calls (Fawcett and Provost, 1997), the detection of oil spills in satellite radar images (Kubat et al., 1998), credit scoring (Brown and Mues, 2012) or student retention (Thammasiri et al., 2014).

In many real-world applications, the class distribution of instances is most often imbalanced and the costs of misclassification are different. Thus, class-imbalance and cost-sensitive learning have attracted much attention from researchers. Sampling is one of the widely used approaches in dealing with the class imbalance problem, which alters the class distribution of instances so that the minority class is well represented in the training data (Thammasiri et al., 2014).

Re-sampling techniques apply a preprocessing step in order to balance the original imbalanced data. In this paper, three balancing techniques were used. Weka's preprocess supervised filter (Witten and Frank, 2005) was used to perform the re-sampling on the dataset. The re-sampling techniques used were (López et al., 2012b): under-sampling, oversampling and mix. A brief description of each of them is given below:

- Random under-sampling: a non-heuristic method that aims to balance class distribution through the random elimination of majority class instances. The elimination of majority class instances is performed in order to try to balance out the data set in an attempt to overcome the idiosyncrasies of the machine learning algorithm. The major drawback of random under-sampling is that this method can discard potentially useful data that could

Table 1
Description of the variables and classification by severity.

Code: Variable	Categories	Description	Count	Severity	
				KSEV	SLIG
VEH: Vehicles involved	1	1 vehicle	3615	875	2740
	2	2 vehicles	11,098	1881	9217
	3	3 vehicles	1711	274	1437
	4	4 vehicles	391	60	331
ACT: Accident type	COL _ANI	Animal	36	12	24
	COL _FOBJ	Fixed Object	3694	852	2842
	COL _FOV	Fall-off-vehicle	52	22	30
	COL _MVEH	Moving vehicle	12,884	2175	10,709
	COL _OT	Other	141	28	113
	COL _PTW	Motorcycle	8	1	7
PAT: Accidentpattern	CIR	Circle	26	4	22
	HEAD	Head-on	11,602	2010	9592
	INT	Intersection	67	4	63
	OVER	Overtake	318	50	268
	REAR	Rear-end	1173	144	1029
	SVEH	Single vehicle	3581	866	2715
	OT	Other	48	12	36
DIR: Directions	1_WAY	One way	2674	455	2219
	2_DIV	Two way divided	9575	1713	7862
	2_UNDIV	Two way undivided	4520	919	3601
	OT	Other	46	3	43
LANE: Number of lanes	1	1 lane	5351	1068	4283
	2	2 lanes	9245	1682	7563
	3	3 lanes	2075	320	1755
	4	4 lanes	144	20	124
TRAME: Horizontal alignment	TG	Tangent	16,660	3054	13,606
	CUR	Curve	155	36	119
GRADE: Road grade	ASC	Ascending	217	37	180
	LEV	Level	16,349	3015	13,334
	DES	Descending	249	38	211
PAV: Pavement type	AS	Asphalt	16,413	3018	13,395
	CON	Concrete	331	61	270
	OT	Dirt, Gravel, Wooden or metal	71	11	60
SUP: Surface condition	IOS	Ice or snow	81	11	70
	DRY	Dry	16,203	2996	13,207
	OT	Mud, sand or oil	21	8	13
	WET	Wet	510	75	435
CATM: Weather conditions	CLE	Clear	16,369	3018	13,351
	RAIN	Rain	233	38	195
	OT	Snow, Storm wind, fog or dust	213	34	179
LIG: Lighting conditions	DARK	Dark	159	42	117
	DAY	Daylight	15,019	2763	12,256
	NIGHT_IN	Night with insufficient lighting	100	17	83
	NIGHT_SU	Night with sufficient lighting	1151	185	966
	OT	Sunrise or sunset	386	83	303
CONT: Traffic control	FTL	Flashing traffic light (red or yellow)	97	19	78
	NO_CONT	No control or non-working traffic sign	16,482	3050	13,432
	OS_Y	Obligatory signs or yield	25	3	22
	PAV_MAR	Pavement markings	3	–	3
	POL	Police officer	62	10	52
	POTS	Police officer with traffic signal	4	–	4
	STOP	Stop	46	1	45
	TSIGN	Traffic signal	96	7	89
SPE: Speed limit	20	20	38	7	31
	30	30	134	16	118
	40	40	4344	604	3740
	50	50	3409	429	2980
	60	60	4804	904	3900
	70	70	1420	330	1090
	80	80	1811	496	1315
	90	90	445	148	297
	100	100	257	95	162
	>=110	>=110	153	61	92

be important for the induction process. In addition, once under-sampling the majority class is performed, the sample can no longer be considered random. This is due to the fact that when using classifiers on some data set, there is no predefined known probability distribution of target population, and since that distribution is unknown, sample distribution is used in attempt to try to estimate the population distribution, and as long as the sample is drawn randomly, the sample distribution can be used to estimate the population distribution from where it was drawn (Kotsiantis et al., 2006).

- Synthetic minority oversampling technique (SMOTE): This is a heuristic method which creates a subset of the original dataset by creating synthetic minority examples; the minority class is over-sampled by taking each minority class sample and introducing synthetic examples along the line of the segments joining any/all of the (k) minority class nearest neighbours. Depending upon the amount of oversampling required, neighbours from the (k) nearest neighbours are randomly chosen and one sample is generated in the direction of each. Synthetic instances are generated in a less application-specific manner, by operating in "variable space" rather than "data space". More specifically, synthetic samples are generated by taking the difference between the variable vector (sample) under consideration and its nearest neighbour and multiplying this difference by a random number between zero and one, and then adding it to the variable vector under consideration. This causes the selection of a random point along the line segment between two specific variables and hence effectively forces the decision region of the minority class to become more general (Chawla et al., 2002).
- Mix method: This method combines both sampling techniques used in under-sampling and oversampling. In this method, the minority class instances are randomly duplicated while randomly discarding the majority class instances in order to modify the class distribution until the number of instances belonging to each class is roughly the same, preserving the original data set size (Witten and Frank, 2005).

2.2. Bayes classifiers

Statistical inference such as classifications in data processing could be performed successfully using the Bayes principle. A Bayes classifier is based on the idea that the role of a class is to predict the values of variables for members of that class, where instances are grouped in classes because they have common values for the variables (Poole and Mackworth, 2010).

2.2.1. Naïve Bayes classifiers

The Naive Bayes (NB) classifier is considered the simplest of Bayes classifiers, which makes the independence assumption that the input variables are conditionally independent of each other given the classification. The NB classifier results could be illustrated using a particular belief network, in which each variable is represented as a node (called child) and the class variable is the only parent for all the other variables.

Let X_i, ($i = 1, 2, \ldots, n$ for n variables) and each variable have values of $x_1, x_2, \ldots, x_n$ that describe each instance. The most probable target value is described as v_{MAP}, while V is a finite set building on every target value v_j. In order to assign the most probable target value of the test instance using Bayes approach for classification, one set of training instances with a specific class is given, a classifier must be learned to predict the class distribution of an instance with its class unknown. The Bayes classifier is defined as (Wu and Cai, 2011):

$$v_{MAP} = \mathrm{argmax}_{v_j \in V} P(v_j) \prod_{i=1}^{n} P(x_1, x_{2,\ldots,} x_n | v_j) \quad (1)$$

NB classifier is best suited when the independence assumption is valid. That is, when the class is a good predictor of the other variables and the other variables are independent given the class, then the NB can be used as follows:

$$v_{NB} = \mathrm{argmax}_{v_j \in V} P(v_j) \prod_{i=1}^{n} P(x_i | v_j) \quad (2)$$

Each variable node in NB has the class node as its parent, but does not have any parent from variable nodes. In many fields including traffic accidents analysis, assuming independence between contributing variables is unrealistic. When two variables are related, NB may place too much weight on the influence from the two variables, and too little on the other variables, which can result in classification bias. However, deleting one of these variables may have the effect of alleviating the problem (Zheng and Webb, 2006).

In order to improve NB's accuracy by weakening its variable independence assumption, Semi-NB classifiers can be used to enhance NB by relieving the restriction of conditional independence amongst variables. One of the techniques used to relieve this restriction are the Aggregate One-Dependence Estimators (AODE), which achieve a higher accuracy by averaging over a constrained group of 1-dependence NB models developed on a small space (Webb et al., 2005).

In AODE, a one-dependence classifier is developed for each variable, where the variable is set to be the parent of all other variables. Then, AODE directly averages the aggregate consisting of many special tree augmented naive Bayes. AODE maintains the simplicity and direct theoretical foundation of NB without incurring the high time, in addition to having good classification performance (Wu and Cai, 2011).

The AODE classifier is defined as follows (Webb et al., 2005):

$$v_{AODE} = \mathrm{argmax}_{v_i \in V} \left(\sum_{i:1 \le i \le n \wedge f(v_i) \ge m} P(x_i, v_i) \prod_{j=1, j \ne i}^{n} P(x_j | x_i, v_j) \right) \quad (3)$$

where $f(v)$ is a count of the number of training instances having variable-value x_i and is used to enforce the limit m that they place on the support needed in order to accept a conditional probability estimate and n is the number of variables.

An improved AODE called AODE with Subsumption Resolution (AODEsr) utilizes the tables of probability estimates formed at training time to efficiently detect and address a special form of dependency between two variable-values at classification time. AODEsr relaxes the variable independence assumption in which each variable depends upon the class and at most one other variable.

The specialization-generalization relationship is one extreme type of interdependence. For two variable values v_i and v_j if $P(x_j|x_i) = 1.0$ then x_j is a generalization of x_i and x_i a specialization of x_j. AODEsr deletes generalization variable-values if a specialization is detected, and aggregates the predictions of all qualified classifiers using the remaining variable-values (Zheng and Webb, 2006). Another semi-NB classifier is the Weightily Averaged One-Dependence Estimators (WAODE), which weighs the averaged 1-dependence classifiers by the conditional mutual information, which significantly outperforms the AODE (Jiang and Zhang, 2006).

WAODE is an extended NB classifier that relaxes the conditional independence assumption of NB, and consists of multiple one-dependence estimators. A One-dependence estimator (ODE) is a classifier with a single variable that is the parent of all other variables. In WAODE, ODEs are constructed for each variable, and a different weight is assigned for each ODE. WAODE averages the aggregate of the weighted ODEs (Jiang and Zhang, 2006).

If an instance E is represented by $E=(x_1, \ldots, x_n)$ where x_i is the value of variable X_i. The WAODE classifier is defined as:

$$v(E) = \text{argmax}_{v \in V} \left(\frac{\sum_{i=1}^{n} W_i P(x_i, v) \prod_{j=1, j \neq i}^{n} P(x_j | x_i, v)}{\sum_{i=1}^{n} W_i} \right) \quad (4)$$

where W_i is the weight of the ODE in which variable X_i is the parent of all other variables. In WAODE, in order to determine the weight W_i, mutual information between variable x_i and class v is used:

$$W_i = \sum_{x_i, v} P(x_i, v) \log \frac{P(x_i, v)}{P(x_i)P(v)} \quad (5)$$

where the probabilities $P(x_i, v)$ and $P(x_j/x_i, v)$ are estimated as follows:

$$P(x_i, v) = \frac{f(x_i, v) + 1.0/(n_i, k)}{N + 1.0} \quad (6)$$

$$P(x_j | x_i, v) = \frac{f(x_j, x_i, v) + 1.0/n_j}{f(x_i, v) + 1.0} \quad (7)$$

where $f(.)$ is the frequency with which a combination of terms appears in the training data, N is the number of training instances, n_i is the number of values of variable x_i, and k is the number of classes.

2.2.2. Bayesian networks classifiers

BNs applications have grown extensively into different fields, with theoretical and computational developments in many areas (Mittal et al., 2007). These include: modelling knowledge in bioinformatics, medicine, document classification, information retrieval, image processing, data fusion, decision support systems, engineering, gaming, and law.

Let $X=\{X_1, \ldots, X_n\}$, $n \geq 1$ be a set of variables. BN over a set of variables X is a network structure, which is a Directed Acyclic Graph over X and a set of probability tables $B_p = \{p(X_i | pa(X_i)), X_i \in X)\}$ where $pa(X_i)$ is the set of parents or antecedents of X_i in BN and $i=(1,2,3, \ldots, n)$. A BN represents joint probability distributions $P(X) = \prod_{X_i \in X} p(x_i | pa(x_i))$.

Relationships between variables based on the theory of BN (Neapolitan, 2009) are represented by arcs in the graph, and could represent causality, relevance or relations of direct dependence between variables. However, for the purpose of this research, we do not assume a causal interpretation of the arcs in the networks such as in Acid et al. (2004). Consequently, the arcs are interpreted as direct dependence relationships between the linked variables, and the absence of arcs means the absence of direct dependence between variables; however, indirect dependence relationships between variables could exist.

The classification task consists in classifying a variable $V=v_i$, called the class variable, given a set of variables $X=X_1, \ldots, X_n$, called attribute variables. A classifier $h: X \rightarrow V$ is a function that maps an instance of X to a value of V. The classifier is learned from a data set D consisting of samples over (X, V). The learning task consists of finding an appropriate BN given a data set D over X.

In order to learn the structure in BNs, two approaches are available: First, the constraint based approach, which performs tests of conditional independence on the data and searches for a network that is consistent with the observed dependences and independences. Second, the score based approach, which defines a score that evaluates how well the dependences or independences in a structure match the data and search for a structure that maximizes the score.

In this study, we used the following Bayes Classifiers: AODE, WAODE and BNs in order to build different models and to compare their results in terms of their ability to correctly classify traffic accidents according to their injury severity into either KSEV or SLIG. When building the models using BNs, three search methods were used: hill climber, hill climber algorithm restricted by an order on the variables (K2) and simulated annealing search algorithm. Also, three different score metrics functions were used: BDe score metric (BDeu); Minimum Description Length (MDL); and the Akaike Information Criterion (AIC).

The search algorithms and the scores were applied in this study mainly because, besides being widely used and being relatively quick, they produce good results in terms of network complexity and accuracy (Madden, 2009). Data sets used, models developed, and their descriptions are presented in Table 2.

2.3. Performance evaluation measures

In order to evaluate the performance of the different developed models, a number of common performance measures were used. These performance measures were calculated using the confusion matrix (see Table 3). In the binary class problem, this matrix shows the results of correctly and incorrectly predicted instances for each class. Where the True Positives (TP) denotes the number of positive (in our case SLIG) instances correctly classified, True Negatives (TN) denotes the number of negative (in our case KSEV) instances correctly classified, False Positives (FP) denotes the number of positive instances incorrectly classified, and False Negatives (FN) denotes the number of negative instances incorrectly classified.

The performance measures used in this study were accuracy, sensitivity, specificity and F-measure. Their equations are:

$$\text{Accuracy} = \frac{\text{TP} + \text{TN}}{\text{TP} + \text{FP} + \text{TN} + \text{FN}} \quad (8)$$

$$\text{Sensitivity} = \frac{\text{TP}}{\text{TP} + \text{FN}} \quad (9)$$

$$\text{Specificity} = \frac{\text{TN}}{\text{TN} + \text{FP}} \quad (10)$$

$$\text{Precision} = \frac{\text{TP}}{\text{TP} + \text{FP}} \quad (11)$$

$$\text{F-measure} = \frac{2 \times \text{Precision} \times \text{Sensitivity}}{\text{Precision} + \text{Sensitivity}} \quad (12)$$

Accuracy is the proportion of instances that were correctly classified among all instances. Accuracy only gives information on the classifier's overall performance. In cases where there is a highly skewed data distribution, the overall accuracy is not sufficient. In this case, accuracy might give a false indication that a classifier performance is high, where in fact the classifier is only predicting all samples as belonging to one class value, in which case it is biased in its results to majority class. Sensitivity and specificity are usually adopted to monitor classification performance on two classes separately. Sensitivity represents the proportion of correctly predicted as SLIG among all the observed as SLIG. Specificity represents the proportion of correctly predicted as KSEV among all the observed KSEV. F-measure represents the harmonic mean of precision and sensitivity, and is frequently used in imbalanced datasets (Wang et al., 2015).

However, a trade-off exists between sensitivity and specificity. Therefore, the area under a Receiver Operating Characteristic (ROC) curve is also used as a target performance measure. ROC curve represents the true positive rate (sensitivity) vs. the false positive rate (1-specificity). ROC curves are more useful as descriptors of overall performance, reflected by the area under the curve, with a maximum of one describing a perfect test and a ROC area of 0.50 describing a valueless test.

For the analysis of traffic accident injury severity and in order to determine the optimal dataset-classifier, the measures described above were first calculated: accuracy, sensitivity, specificity,

Table 2
Data sets used and developed models description.

Dataset	Classifier	Search	Score	Code
Original dataset (OD)	AODEsr	–	–	OD-AODEsr[a]
	WAODE	–	–	OD-WAODE[b]
	BayesNet	HILLCLIBER	BDeu	OD-Bayes[c]. Hill.Bdeu
	BayesNet	HILLCLIBER	MDL	OD-Bayes.Hill.MDL
	BayesNet	HILLCLIBER	AIC	OD-Bayes.Hill.AIC
	BayesNet	K2	BDeu	OD-Bayes.K2.Bdeu
	BayesNet	K2	MDL	OD-Bayes.K2.MDL
	BayesNet	K2	AIC	OD-Bayes.K2.AIC
	BayesNet	SimulatedAnnealing	BDeu	OD-Bayes.Simulated.Bdeu
	BayesNet	SimulatedAnnealing	MDL	OD-Bayes.Simulated.MDL
	BayesNet	SimulatedAnnealing	AIC	OD-Bayes.Simulated.AIC
Under-sample (US)	AODEsr	–	–	US-AODEsr
	WAODE	–	–	US-WAODE
	BayesNet	HILLCLIBER	BDeu	US-Bayes.Hill.Bdeu
	BayesNet	HILLCLIBER	MDL	US-Bayes.Hill.MDL
	BayesNet	HILLCLIBER	AIC	US-Bayes.Hill.AIC
	BayesNet	K2	BDeu	US-Bayes.K2.Bdeu
	BayesNet	K2	MDL	US-Bayes.K2.MDL
	BayesNet	K2	AIC	US-Bayes.K2.AIC
	BayesNet	SimulatedAnnealing	BDeu	US-Bayes.Simulated.Bdeu
	BayesNet	SimulatedAnnealing	MDL	US-Bayes.Simulated.MDL
	BayesNet	SimulatedAnnealing	AIC	US-Bayes.Simulated.AIC
Oversample (OS)	AODEsr	–	–	OS-AODEsr
	WAODE	–	–	OS-WAODE
	BayesNet	HILLCLIBER	BDeu	OS-Bayes.Hill.Bdeu
	BayesNet	HILLCLIBER	MDL	OS-Bayes.Hill.MDL
	BayesNet	HILLCLIBER	AIC	OS-Bayes.Hill.AIC
	BayesNet	K2	BDeu	OS-Bayes.K2.Bdeu
	BayesNet	K2	MDL	OS-Bayes.K2.MDL
	BayesNet	K2	AIC	OS-Bayes.K2.AIC
	BayesNet	SimulatedAnnealing	BDeu	OS-Bayes.Simulated.Bdeu
	BayesNet	SimulatedAnnealing	MDL	OS-Bayes.Simulated.MDL
	BayesNet	SimulatedAnnealing	AIC	OS-Bayes.Simulated.AIC
Mix (MS)	AODEsr	–	–	MS-AODEsr
	WAODE	–	–	MS-WAODE
	BayesNet	HILLCLIBER	BDeu	MS-Bayes.Hill.Bdeu
	BayesNet	HILLCLIBER	MDL	MS-Bayes.Hill.MDL
	BayesNet	HILLCLIBER	AIC	MS-Bayes.Hill.AIC
	BayesNet	K2	BDeu	MS-Bayes.K2.Bdeu
	BayesNet	K2	MDL	MS-Bayes.K2.MDL
	BayesNet	K2	AIC	MS-Bayes.K2.AIC
	BayesNet	SimulatedAnnealing	BDeu	MS-Bayes.Simulated.Bdeu
	BayesNet	SimulatedAnnealing	MDL	MS-Bayes.Simulated.MDL
	BayesNet	SimulatedAnnealing	AIC	MS-Bayes.Simulated.AIC

[a] Averaged One-Dependence Estimators.
[b] Weightily Average One-Dependence Estimators.
[c] Bayes network.

F-measure and ROC area. Later, the best Bayes classifier model found in terms of these measures was used for analysis.

2.4. Bayes network inference

Inference in BNs consists of computing the conditional probability of some variables, given that other variables are set to evidence. Inference may be done for a specific state or value of a variable, given evidence on the state of other variable(s). Thus, using the conditional probability table for the BN developed; their values can be easily inferred. See De Oña et al. (2011) for a detailed explanation and examples. When using BNs, inference is necessary to interpret the results from a road safety perspective.

Table 3
Confusion matrix for a binary class problem.

		Predicted class	
		Positive	Negative
Actual class	Positive	True positive (TP)	False positive (FP)
	Negative	False negative (FN)	True negative (TN)

3. Results

In this section, an analysis was performed to determine the performance of the different alternatives used for the imbalanced data set classification. The aim was to analyze three different issues:

- The improvement obtained by resampling data sets using three different resample techniques: under-sampling, oversampling and mix versus the original imbalanced data set.
- The possible resulting differences between balanced and imbalanced data sets as measured by applying the different Bayes classifiers and their resulting performance evaluation measures.
- The risk factors that are significantly associated with the occurrence of a killed or a severe injury in a traffic accident were extracted using the classifier that presented the best performance in terms of evaluation measures.

First, the experimental framework with the data set employed in the analysis is presented and then, the performance of the different classifiers that will allow supporting the extracted findings is

Table 4
Number of accidents and severity distribution in the different datasets.

Dataset	Total	SLIG	KSEV
Original dataset	16,815	13,725	3090
Under-sampling	6180	3090	3090
Oversampling	27,450	13,725	13,725
Mix	16,815	8414	8401

shown. Finally, the main issues that arise from the aforementioned analysis are discussed.

3.1. Imbalanced versus balanced data sets

The original dataset contained 16,815 accidents in which the injury severity distribution was: 13,725 slight injuries and 3090 killed or severe injuries, and in which the target variable (severity) was predominantly imbalanced. To deal with the imbalanced data set problem, three new balanced data sets were developed using three different resample techniques: under-sampling, oversampling and mix. Table 4 shows the total number of records (instances) in all the datasets used and their distribution amongst different severity classes.

As shown in Table 4, when the random under-sampling technique was used, the dataset was reduced to the size of the minority class, in this case to killed or severe class (3090 instances for slight class as shown in Table 1). While when using SMOTE oversampling, the number of the instances in the resulting dataset was increased to the size of the majority class (13,725 instances for killed or severe class). Finally, in the mix sample, the resulting dataset preserved the original number of instances (16,815 accidents), however, the majority class instances were reduced to 8414 instances and the minority class was increased to 8401 instances.

3.2. Bayes classifiers and data sets

The different Bayes classifiers described in Section 2 were used to build different models. For each dataset (original, oversample, under-sample and mix), eleven models were developed. Each model was first divided into ten subsamples, nine were used to train the model and the last was used to test the model. The process was repeated ten times (runs) for each subsample (ten-fold cross-validation). The description of all the models developed is shown in Table 2.

The results of the models developed for each balanced dataset were then compared with those obtained using the original dataset models. In order to perform this comparison, the results of the evaluation measures used to compare the models developed are summarized in Table 5. The comparison is based on the performance measures of accuracy, sensitivity, specificity, ROC Area and F-measure (the results shown are the averages of the ten runs).

Table 5 shows the average results of the performance measures used for the different models developed, where a corrected paired *t*-test was used to test their statistical significance. To that end, forty-four models were developed using a combination of datasets with different classifiers. With respect to the results obtained by the testing set, the following findings were extracted:

- None of the balanced datasets showed a statistically significant improvement when compared to the original dataset based on accuracy, sensitivity, and F-measure. This result however was expected, since these measures are highly biased to the majority class (SLIG) as seen in their equations presented in Section 2.3.
- The results obtained by specificity showed that the original dataset had the worst results in which the models used were completely incapable of classifying KSEV correctly. Results also showed that the highest result obtained using this measure on the original dataset was when using k2.AIC with a result of 0.030 which shows a drastic incapability of classification. On the other hand, most of the balanced datasets showed a statistically significant improvement when compared to the original dataset, where the highest result obtained was when using the following dataset/model respectively: Oversample/k2.AIC with a result of 0.650.
- ROC Area indicated that the following model/dataset showed the highest significant statistical improvement when compared to the original dataset: Oversample/Hillclimber.AIC with a result of 0.680, Oversample/k2.AIC with a result of 0.680, Oversample/Simulated.AIC with a result of 0.680.
- The total number of wins shown in the last row in Table 5 represents the total number of the times the dataset used had a statistically significant improvement when compared to the original dataset. As seen, the oversample dataset had twenty-two wins indicating the best performance when compared to the under-sample dataset having a total of eleven wins and the mix dataset having a total of sixteen wins.

Based on the aforementioned results, balanced datasets performed better than the original in terms of its ability to classify the minority class (KSEV), with the best models obtained using the oversample dataset.

In addition, performance of the models developed using BNs was superior to those developed using other classifiers; consequently, BNs performance was compared to WAODE and AODEsr. Table 6 shows the comparison of two models developed with Bayes classifiers (AODErs and WAODE) versus the rest of the BNs in both oversample and original data sets.

Based on Table 6, the following results were extracted:

- The best performing classifiers that obtained a statistically significant improvement in terms of accuracy in the original data set were Bayes.Hill.Bdeu and Bayes.Hill.MDL where they both scored 0.816 when compared to AODEsr and WAODE, and where in the oversample data set Bayes.Simulated.AIC and WAODE had values of 0.628 and 0.627 respectively when compared to AODEsr. This result is due to the fact that in imbalanced data sets, accuracy is not a proper measure since it does not distinguish between the numbers of correctly classified instances of different classes and may lead to erroneous conclusions. López et al. (2013) stated that a classifier achieving an accuracy of 90% in a dataset with an imbalanced data set is not accurate if it classifies most examples as belonging to the class with more instances.
- In terms of sensitivity, again the best performing classifiers in the original data set were Bayes.Hill.Bdeu and Bayes.Hill.MDL where they both had a value of one when compared to AODEsr and WAODE, while in the oversample data set; AODEsr had a value of 0.663 when compared to WAODE. Since sensitivity measures the capability of the classifier to classify SLIG, and the original data set is biased to SLIG, therefore, the performance obtained reflected this fact.
- Specificity results for the original data set indicated that the best performance was obtained using Bayes.k2.AIC, with a value 0.029 when compared to AODEsr and WAODE. On the other hand, in the oversample data set, the best classifier was the Bayes.k2. AIC, with a value of 0.655 when compared to both AODEsr and WAODE. Here it is evident that none of the used classifiers in the original data set was capable of classifying KSEV correctly, while in the balanced data set, all classifiers showed a remarkable improved capability of classifying KSEV.
- Regarding F-measure, the results indicated that both Bayes.Hill.Bdeu and Bayes.Hill.MDL had a statistically significant improvement when compared to both AODEsr and

Table 5
Comparison between the different models using performance measure and paired *t*-test.

Dataset	Dataset		Datasets compared to base					
	Original		Under-sample		Oversample		Mix	
	Measure value	s.d ±	Measure value	s.d ±	Measure value	s.d ±	Measure value	s.d ±
Performance measure	*Accuracy*							
AODEsr	0.815	0.016	0.595	0.169	0.623	0.087	0.611	0.119
WAODE	0.814	0.018	0.599	0.177	0.627	0.081	0.619	0.123
Hillclimber.Bdeu	0.816	0.002	0.592	0.169	0.619	0.084	0.583	0.130
Hillclimber.MDL	0.816	0.004	0.591	0.173	0.616	0.089	0.585	0.124
Hillclimber.AIC	0.816	0.017	0.592	0.175	0.623	0.091	0.613	0.114
k2.Bdeu	0.815	0.016	0.592	0.168	0.618	0.077	0.601	0.116
k2.MDL	0.815	0.016	0.564	0.292	0.619	0.080	0.601	0.116
k2.AIC	0.815	0.021	0.595	0.179	0.623	0.082	0.615	0.127
Simulated.Bdeu	0.815	0.017	0.591	0.173	0.619	0.079	0.603	0.114
Simulated.MDL	0.815	0.017	0.583	0.261	0.614	0.079	0.604	0.114
Simulated.AIC	0.815	0.020	0.594	0.166	0.628	0.088	0.613	0.126
Performance measure	*Sensitivity*							
AODEsr	1.00	0.00	0.66	0.03	0.66	0.02	0.67	0.02
WAODE	1.00	0.00	0.65	0.03	0.65	0.01	0.66	0.02
Hillclimber.Bdeu	1.00	0.00	0.68	0.04	0.62	0.03	0.75	0.07
Hillclimber.MDL	1.00	0.00	0.68	0.05	0.62	0.03	0.75	0.07
Hillclimber.AIC	1.00	0.00	0.66	0.04	0.61	0.04	0.66	0.03
k2.Bdeu	1.00	0.00	0.68	0.05	0.64	0.01	0.63	0.02
k2.MDL	1.00	0.00	0.73	0.07	0.64	0.01	0.63	0.02
k2.AIC	0.99	0.00	0.65	0.04	0.59	0.03	0.65	0.02
Simulated.Bdeu	1.00	0.00	0.68	0.05	0.64	0.01	0.64	0.02
Simulated.MDL	1.00	0.00	0.70	0.06	0.64	0.01	0.65	0.02
Simulated.AIC	0.99	0.00	0.67	0.04	0.65	0.01	0.64	0.02
Performance measure	*Specificity*							
AODEsr	0.01	0.00	0.53*	0.03	0.58*	0.02	0.55*	0.02
WAODE	0.01	0.00	**0.55***	0.03	0.60*	0.01	**0.58***	0.02
Hillclimber.Bdeu	0.00	0.00	0.50*	0.04	0.62*	0.03	0.41*	0.07
Hillclimber.MDL	0.00	0.00	0.50*	0.05	0.61*	0.03	0.42*	0.07
Hillclimber.AIC	0.01	0.01	0.52*	0.04	0.64*	0.03	0.57*	0.03
k2.Bdeu	0.01	0.01	0.51*	0.04	0.60*	0.02	**0.58***	0.02
k2.MDL	0.01	0.01	0.40*	0.11	0.60*	0.01	**0.58***	0.02
k2.AIC	0.03	0.01	0.54*	0.04	**0.65***	0.03	**0.58***	0.02
Simulated.Bdeu	0.01	0.01	0.51*	0.04	0.60*	0.02	0.56*	0.02
Simulated.MDL	0.01	0.01	0.47*	0.09	0.59*	0.01	0.56*	0.02
Simulated.AIC	0.02	0.01	0.52*	0.04	0.61*	0.02	**0.58***	0.02
Performance measure	*F-measure*							
AODEsr	0.90	0.00	0.62	0.02	0.64	0.01	0.63	0.01
WAODE	0.90	0.00	0.62	0.02	0.64	0.01	0.63	0.01
Hillclimber.Bdeu	0.90	0.00	0.63	0.02	0.62	0.01	0.64	0.03
Hillclimber.MDL	0.90	0.00	0.62	0.02	0.62	0.02	0.64	0.03
Hillclimber.AIC	0.90	0.00	0.62	0.02	0.62	0.02	0.63	0.01
k2.Bdeu	0.90	0.00	0.62	0.02	0.63	0.01	0.61	0.01
k2.MDL	0.90	0.00	0.63	0.02	0.63	0.01	0.61	0.01
k2.AIC	0.90	0.00	0.62	0.02	0.61	0.02	0.63	0.01
Simulated.Bdeu	0.90	0.00	0.62	0.02	0.63	0.01	0.62	0.01
Simulated.MDL	0.90	0.00	0.63	0.02	0.62	0.01	0.62	0.01
Simulated.AIC	0.90	0.00	0.62	0.02	0.63	0.01	0.63	0.01
Performance measure	*ROC area*							
AODEsr	0.63	0.02	0.63	0.02	0.67*	0.01	0.65*	0.01
WAODE	0.64	0.02	0.63	0.02	0.67*	0.01	**0.66***	0.01
Hillclimber.Bdeu	0.61	0.02	0.62	0.02	0.66*	0.01	0.61	0.02
Hillclimber.MDL	0.61	0.02	0.62	0.02	0.66*	0.01	0.62	0.02
Hillclimber.AIC	0.62	0.02	0.63	0.02	**0.68***	0.01	0.65*	0.01
k2.Bdeu	0.63	0.02	0.63	0.02	0.67*	0.01	0.63	0.01
k2.MDL	0.63	0.02	0.58	0.05	0.66*	0.01	0.63	0.01
k2.AIC	0.63	0.02	0.63	0.02	**0.68***	0.01	**0.66***	0.01
Simulated.Bdeu	0.63	0.02	0.63	0.02	0.67*	0.01	0.63	0.01
Simulated.MDL	0.63	0.02	0.61	0.04	0.66*	0.01	0.63	0.01
Simulated.AIC	0.63	0.02	0.63	0.02	**0.68***	0.01	0.65*	0.01
Total number of wins			11		22		16	

The bold numbers indicate that the values obtained are the highest for the respective variable and/or category of variable when compared to others.
s.d: Standard deviation.
* Statistically significant when tested against original dataset using corrected paired *t*-test.

Table 6
Comparison between the performance measures in the models built using oversample and original data set.

Data set		Oversampled data set				Original data set			
Performance measure	Bayes classifier	Classifier compared				Classifier compared			
		AODEsr		WAODE		AODEsr		WAODE	
		Measure value	s.d	Measure value	s.d	Measure value	s.d	Measure value	s.d
Accuracy	*AODEsr*	–	–	0.623	0.087	–	–	0.815	0.016
	WAODE	**0.627***	0.081	–	–	0.814	0.018	–	–
	Bayes.Hill.Bdeu	0.619	0.084	0.619	0.084	**0.816***	0.003	**0.816***	0.003
	Bayes.Hill.MDL	0.616	0.089	0.616	0.089	**0.816***	0.004	**0.816***	0.004
	Bayes.Hill.AIC	0.623	0.091	0.623	0.091	0.816	0.017	0.816	0.017
	Bayes.K2.Bdeu	0.618	0.077	0.618	0.077	0.815	0.016	0.815	0.016
	Bayes.K2.MDL	0.619	0.080	0.619	0.080	0.815	0.016	0.815	0.016
	Bayes.K2.AIC	0.623	0.082	0.623	0.082	0.815	0.021	0.815	0.021
	Bayes.Simulated.Bdeu	0.619	0.079	0.619	0.079	0.815	0.017	0.815	0.017
	Bayes.Simulated.MDL	0.614	0.079	0.614	0.079	0.815	0.017	0.815	0.017
	Bayes.Simulated.AIC	**0.628***	0.088	0.628	0.088	0.815	0.020	0.815	0.020
Sensitivity	*AODEsr*	–	–	**0.663***	0.016	–	–	0.997	0.002
	WAODE	0.654	0.011	–	–	0.996	0.002	–	–
	Bayes.Hill.Bdeu	0.619	0.03	0.619	0.03	**1.000***	0.000	**1.000***	0.000
	Bayes.Hill.MDL	0.625	0.031	0.625	0.031	**1.000***	0.000	**1.000***	0.000
	Bayes.Hill.AIC	0.607	0.039	0.607	0.039	0.998	0.003	0.998	0.003
	Bayes.K2.Bdeu	0.637	0.014	0.637	0.014	0.996	0.002	0.996	0.002
	Bayes.K2.MDL	0.641	0.013	0.641	0.013	0.996	0.002	0.996	0.002
	Bayes.K2.AIC	0.591	0.033	0.591	0.033	0.992	0.003	0.992	0.003
	Bayes.Simulated.Bdeu	0.636	0.015	0.636	0.015	0.996	0.002	0.996	0.002
	Bayes.Simulated.MDL	0.638	0.014	0.638	0.014	0.996	0.002	0.996	0.002
	Bayes.Simulated.AIC	0.645	0.014	0.645	0.014	0.993	0.003	0.993	0.003
Specificity	*AODEsr*	–	–	0.583	0.019	–	–	0.006	0.004
	WAODE	0.600*	0.013	–	–	0.006	0.005	–	–
	Bayes.Hill.Bdeu	0.619*	0.026	0.619*	0.026	0.000	0.001	0.000	0.001
	Bayes.Hill.MDL	0.606*	0.028	0.606	0.028	0.000	0.001	0.000	0.001
	Bayes.Hill.AIC	0.639*	0.033	0.639*	0.033	0.007	0.009	0.007	0.009
	Bayes.K2.Bdeu	0.600*	0.016	0.600	0.016	0.011*	0.006	0.011*	0.006
	Bayes.K2.MDL	0.597*	0.014	0.597	0.014	0.011*	0.006	0.011*	0.006
	Bayes.K2.AIC	**0.655***	0.026	**0.655***	0.026	**0.029***	0.012	**0.029***	0.012
	Bayes.Simulated.Bdeu	0.602*	0.016	0.602	0.016	0.011*	0.006	0.011*	0.006
	Bayes.Simulated.MDL	0.589	0.014	0.589	0.014	0.011*	0.006	0.011*	0.006
	Bayes.Simulated.AIC	0.611*	0.016	0.611*	0.016	0.023*	0.013	0.023*	0.013
F-measure	*AODEsr*	–	–	0.637	0.009	–	–	0.898	0.001
	WAODE	0.637	0.008	–	–	0.897	0.001	–	–
	Bayes.Hill.Bdeu	0.619	0.015	0.619	0.015	**0.899***	0.000	**0.899***	0.000
	Bayes.Hill.MDL	0.619	0.015	0.619	0.015	**0.899***	0.000	**0.899***	0.000
	Bayes.Hill.AIC	0.616	0.019	0.616	0.019	0.898	0.001	0.898	0.001
	Bayes.K2.Bdeu	0.625	0.009	0.625	0.009	0.898	0.001	0.898	0.001
	Bayes.K2.MDL	0.627	0.008	0.627	0.008	0.898	0.001	0.898	0.001
	Bayes.K2.AIC	0.610	0.017	0.610	0.017	0.898	0.001	0.898	0.001
	Bayes.Simulated.Bdeu	0.625	0.009	0.625	0.009	0.898	0.001	0.898	0.001
	Bayes.Simulated.MDL	0.623	0.009	0.623	0.009	0.898	0.001	0.898	0.001
	Bayes.Simulated.AIC	0.634	0.009	0.634	0.009	0.898	0.001	0.898	0.001
ROC area	*AODEsr*	–	–	0.668	0.009	–	–	0.634	0.017
	WAODE	0.671*	0.009	–	–	0.636	0.016	–	–
	Bayes.Hill.Bdeu	0.663	0.009	0.663	0.009	0.608	0.017	0.608	0.017
	Bayes.Hill.MDL	0.658	0.009	0.658	0.009	0.609	0.016	0.609	0.016
	Bayes.Hill.AIC	0.676*	0.009	0.676	0.009	0.623	0.016	0.623	0.016
	Bayes.K2.Bdeu	0.666	0.008	0.666	0.008	0.632	0.017	0.632	0.017
	Bayes.K2.MDL	0.663	0.009	0.663	0.009	0.632	0.017	0.632	0.017
	Bayes.K2.AIC	0.677*	0.008	0.677*	0.008	0.631	0.017	0.631	0.017
	Bayes.Simulated.Bdeu	0.667	0.008	0.667	0.008	0.631	0.016	0.631	0.016
	Bayes.Simulated.MDL	0.659	0.009	0.659	0.009	0.634	0.017	0.634	0.017
	Bayes.Simulated.AIC	**0.680***	0.009	**0.680***	0.009	0.63	0.017	0.630	0.017

The bold numbers indicate that the values obtained are the highest for the respective variable and/or category of variable when compared to others.
s.d: Standard deviation.
* Statistically significant when tested against either AODEsr or WAODE using corrected paired *t*-test.

WAODE in the original data set with a value of 0.899. On the contrary, none of the classifiers in the oversample data set showed any statistically significant improvement. According to Wang et al. (2015), F-measure is frequently used in imbalanced datasets, however, Daskalaski et al. (2006) indicated that higher F-measure means that the model performs better on positive class (i.e. SLIG).

- A well-known approach to unify these measures is the ROC area, which provides a single measure of a classifier's performance for evaluating which model is better on average (López et al., 2013). Herein, more than one BN model obtained high ROC area value with statistically significant improvement where the highest values were obtained for Bayes.Hill.AIC, Bayes.k2.AIC, and Bayes.Simulated.AIC with values of 0.676, 0.677, and 0.680

Table 7
Total number of wins per classifier and measure using oversampled and original data sets.

Data set	Oversample		Original	
Classifier	Total no. of wins	Shown in measures	Total no. of wins	Shown in measures
AODEsr	1	Sensitivity	0	–
WAODE	3	Accuracy Specificity ROC area	0	–
Bayes.Hill.Bdeu	2	Specificity	6	Accuracy Sensitivity F-measure
Bayes.Hill.MDL	1	Specificity	6	Accuracy Sensitivity F-measure
Bayes.Hill.AIC	3	Specificity ROC area	0	–
Bayes.K2.Bdeu	1	Specificity	2	Specificity
Bayes.K2.MDL	1	Specificity	2	Specificity
Bayes.K2.AIC	4	Specificity ROC area	2	Specificity
Bayes.Simulated.Bdeu	1	Specificity	2	Specificity
Bayes.Simulated.MDL	0	–	2	Specificity
Bayes.Simulated.AIC	5	Accuracy Specificity ROC area	2	Specificity

respectively in the oversample data set. On the contrary, none of the classifiers in the original data set had any statistically significant improvement.

Based on the results obtained in Table 6 and the discussion above, none of the classifiers in the original data set were capable of correctly classifying KSEV. However, all the oversample's BNs presented an improved performance, indicating their ability to classify KSEV better.

In order to better illustrate the results obtained, Table 7 shows the total number of times each classifier had statistically significant improvement when compared to either AODEsr or WAODE (total no. of wins) in both the original and the oversample data sets.

As shown in Table 7, the classifiers in the original data set that had the highest number of wins were Bayes.Hill.BDeu and Bayes.Hill.MDL, however, their wins were in the evaluators whose results are sensitive to imbalanced data sets (accuracy, sensitivity, and F-measure). On the other hand, the classifiers in the oversample data set that had the highest number of wins were Bayes.K2.AIC and Bayes.Simulated.AIC, where they both had an improvement in specificity and ROC area, which means that their capability of correctly classifying KSEV (as measured using specificity) and their overall performance (as measured using ROC area) were considerably higher than the models developed using the original data set. Thus, using the oversample data set in combination with Bayes.K2.AIC or Bayes.Simulated.AIC obtained improved results in comparison with the original data set.

3.3. *Injury severity analysis*

Based on the aforementioned results, BNs were used in order to identify the main factors that contribute to classifying an accident according to a specific injury severity. As shown in Table 8 and based on BNs developed using Bayes.Simulated.AIC and Bayes.K2.AIC in both original and oversample datasets, the complexity (number of arcs) of the developed BNs was calculated. In addition, the direct dependence relationships between severity (SEV) and the rest of the variables as well as interdependences amongst the different variables were illustrated. Furthermore, relationships were grouped so that they take into account whether or not the relation was directly related with severity.

As illustrated in Table 8, only three variables had direct dependence with SEV in the original data set in Bayes.k2.AIC, whereas only two variables in Bayes.k2.AIC had direct dependence with SEV, where the only common variable between the two models was speed (SPE). In the oversample data set, eleven variables had a direct dependence relationship with SEV in the Bayes.Simultaned.AIC models, whereas in the model Bayes.K2.AIC all the thirteen variables presented this direct relation with SEV.

Thus, variables that were found to be directly related to SEV in both models were: vehicles involved (VEH), accident type (ACT), accident pattern (PAT), direction (DIR), number of lanes (LANE), grade (GRADE), pavement type (PAV), surface condition (SUP), speed (SPE), traffic control (CONT) and lighting conditions (LIG). The fact that these variables appeared in the two models indicates that these variables have a strong dependence on SEV.

Many researchers also found that these variables were of the influencing factors that affect injury severity, such as Manner and Wünsch-Ziegler (2013) and Kadilar (in press) who found that road condition and speed limit both affected injury severity. Accident type was also found to affect injury severity by De Oña et al. (2011), Theofilatos et al. (2012), Kadilar (in press) and most recently by Manner and Wünsch-Ziegler (2013). According to Kadilar (in press) and Kockelman and Kweon (2002) the number of vehicles contributes significantly to the resulting injury severity, while lighting was found to be significant by both Ma et al. (2009) and De Oña et al. (2011). Ma et al. (2009) found that, amongst other variables, road alignment was one of the contributing variables that affect severity while Theofilatos et al. (2012) showed that time of accident had also a significant effect.

On the other hand, in the original data set, Bayes.k2.AIC had a complexity of twenty-three with only one new interdependence, which was between horizontal alignment (TRAME) and LANE; whereas Bayes.Simulated.AIC had a complexity of twenty-one with three new interdependences between horizontal alignment (TRAME) with DIR, and SUP with both LIG, SEV.

Regarding complexity in the oversample data set, Bayes.K2.AIC had more arrows than all the other models (thirty-two arrows) with two new dependences between variables not found in the other models, which were the dependences between SEV with both TRAME and weather conditions (CATM). Bayes.Simulated.AIC had thirty-one arrows with seven new interdependences, which were between the following variables: DIR with PAV; PAT with LIG; ACT

Table 8
Relationships between variables in the BNs models built using oversampled and original data set.

Data set	Oversampled						Original					
Classifier	Bayes.K2.AIC			Bayes.Simulated.AIC			Bayes.K2.AIC			Bayes.Simulated.AIC		
Dependence relationships between SEV and others variables	SEV	->	VEH	SEV	->	VEH	SEV	->	VEH	–	–	–
	SEV	->	ACT	SEV	->	ACT	–	–	–	SEV	->	ACT
	SEV	->	PAT	SEV	->	PAT	–	–	–	–	–	–
	SEV	->	DIR	SEV	->	DIR	SEV	->	DIR	–	–	–
	SEV	->	LANE	SEV	->	LANE	–	–	–	–	–	–
	SEV	->	GRADE	SEV	->	GRADE	–	–	–	–	–	–
	SEV	->	PAV	SEV	->	PAV	–	–	–	–	–	–
	SEV	->	SUP	SEV	->	SUP	–	–	–	–	–	–
	SEV	->	LIG	SEV	->	LIG	–	–	–	–	–	–
	SEV	->	SPE	SEV	->	SPE	SEV	->	SPE	SEV	->	SPE
	SEV	->	CONT	SEV	->	CONT	–	–	–	–	–	–
	SEV	->	TRAME	–	–	–	–	–	–	–	–	–
	SEV	->	CATM	–	–	–	–	–	–	–	–	–
Interdependence relationships between different variables	VEH	->	ACT	–	–	–	VEH	->	ACT	–	–	–
	VEH	->	PAT	–	–	–	VEH	->	PAT	–	–	–
	VEH	->	DIR	VEH	->	DIR	VEH	->	DIR	VEH	->	DIR
	VEH	->	LANE	–	–	–	VEH	->	LANE	–	–	–
	DIR	->	LANE	DIR	->	LANE	DIR	->	LANE	DIR	->	LANE
	DIR	->	TRAME	–	–	–	DIR	->	TRAME	–	–	–
	DIR	->	LIG	–	–	–	DIR	->	LIG	DIR	->	LIG
	DIR	->	SPE	DIR	->	SPE	DIR	->	SPE	DIR	->	SPE
	DIR	->	CONT	–	–	–	DIR	->	CONT	DIR	->	CONT
	–	–	–	DIR	->	PAV	–	–	–	–	–	–
	–	–	–	–	–	–	–	–	–	–	–	–
	GRADE	->	PAV	–	–	–	GRADE	->	PAV	–	–	–
	GRADE	->	SUP	–	–	–	GRADE	->	SUP	GRADE	->	SUP
	GRADE	->	CATM	GRADE	->	CATM	GRADE	->	CATM	GRADE	->	CATM
	–	–	–	GRADE	->	TRAME	–	–	–	GRADE	->	TRAME
	LANE	->	PAV	–	–	–	LANE	->	PAV	–	–	–
	LANE	->	SPE	–	–	–	LANE	->	SPE	–	–	–
	–	–	–	–	–	–	LANE	->	TRAME	–	–	–
	TRAME	->	GRADE	–	–	–	TRAME	->	GRADE	–	–	–
	–	–	–	–	–	–	–	–	–	TRAME	->	DIR
	PAV	->	SUP	PAV	->	SUP	PAV	->	SUP	PAV	->	SUP
	–	–	–	PAV	->	GRADE	–	–	–	PAV	->	GRADE
	–	–	–	PAV	->	CATM	–	–	–	PAV	->	CATM
	SUP	->	CATM	–	–	–	SUP	->	CATM	–	–	–
	–	–	–	SUP	->	TRAME	–	–	–	SUP	->	TRAME
	–	–	–	–	–	–	–	–	–	SUP	->	LIG
	–	–	–	–	–	–	–	–	–	SUP	->	SEV
	CATM	->	LIG	–	–	–	CATM	->	LIG	–	–	–
	–	–	–	CATM	->	SUP	–	–	–	CATM	->	SUP
	PAT	->	GRADE	–	–	–	PAT	->	GRADE	–	–	–
	–	–	–	PAT	->	LIG	–	–	–			
	–	–	–	PAT	->	VEH	–	–	–	PAT	->	VEH
	–	–	–	ACT	->	PAT	–	–	–	ACT	->	PAT
	–	–	–	ACT	->	SPE	–	–	–	–	–	–
	–	–	–	SPE	->	LANE	–	–	–	SPE	->	LANE
	–	–	–	SPE	->	CONT	–	–	–	–	–	–
	–	–	–	LIG	->	DIR	–	–	–	–	–	–
	–	–	–	LIG	->	GRADE	–	–	–	–	–	–
	–	–	–	LIG	->	CATM	–	–	–	–	–	–
New dependece[a]	2			7			1			3		
Complexity[b]	32			31			23			21		

Note: -> indicates direct dependence relationships.
[a] This row shows number of new dependence relationships found in a classifier that were not found in others.
[b] This row shows the total number of dependence relationships obtained using each classifier.

with SPE; SPE with CONT; LIG with DIR; LIG with GRADE and finally LIG with CATM.

3.4. Bayesian networks inference

Inference was used to determine the most significant variables that were associated with the occurrence of KSEV in traffic accidents. Table 9 presents values of variables that contributed the most to the occurrence of a KSEV outcome in a traffic accident. For each variable, the probability of a value was set to be one (setting evidence) and the other values of the same variable were set to be zero. Consequently, the associated probabilities of severity were calculated.

Table 9 shows the values of variables in which the probability of a KSEV was found to be higher than that of a SLIG. For example, this table shows that assigning a probability of one to the value VEH = 1 (only one vehicle was involved in a traffic accident), results in a probability of KSEV of 0.5875 in the K2.AIC model and 0.5877 in the Simulated.AIC model when using the oversample data set.

These probabilities were calculated from the conditional probability table of the BN developed, and since it was intended to

Table 9
Inference results for variables that are associated with KSEV in traffic accidents in BNs built using oversampled and original data set.

Data set	Oversampled		Original	
Variable and its evident value associated with SEV	Probability of KSEV in K2.AIC model	Probability of KSEV in Simulated.AIC model	Probability of KSEV in K2.AIC model	Probability of KSEV in Simulated.AIC model
VEH = 1	**0.5875**	**0.5877**	0.2420	0.2328
PAT = SVEH	**0.589**	**0.5892**	0.2420	0.2332
GRADE = LEV	0.5049	0.5049	0.1837	0.1840
DIR = 2_DIV	0.5043	0.5042	0.1787	0.1847
DIR = 2_UNDIV	**0.5262**	**0.5259**	0.2033	0.1822
CONT = NO_CONT	0.5040	0.5039	0.1838	0.1839
ACT = COL_FOV	**0.7576**	**0.7625**	0.2384	0.4246
ACT = COL_FOBJ	**0.5729**	**0.5728**	0.2360	0.2307
ACT = COL_ANI	**0.6579**	**0.6643**	0.2308	0.3379
LANE = 1	0.5254	0.5254	0.1878	0.1778
TRAME = TG	0.5008	0.501	0.1837	0.1838
PAV = ASP	0.5042	0.5042	0.1837	0.1841
LIG = DAY	0.5083	0.5084	0.1837	0.1841
LIG = DARK	**0.595**	**0.597**	0.1880	0.1848
SUP = DRY	0.5046	0.5035	0.1837	0.1849
SUP = OT	**0.6188**	**0.6461**	0.1842	0.3863
CATM = CLE	0.5039	0.5015	0.1837	0.1843
SPE = 60	0.5092	0.5107	0.1872	0.1890
SPE = 70	**0.5712**	**0.5707**	0.2344	0.2322
SPE = 80	**0.6239**	**0.624**	0.2737	0.2757
SPE = 90	**0.6867**	**0.6871**	0.3356	0.3350
SPE = 100	**0.7188**	**0.7197**	0.3708	0.3761
SPE = 110	**0.7424**	**0.7424**	0.4075	0.4174

The bold numbers indicate that the values obtained are the highest for the respective variable and/or category of variable when compared to others.

determine which values of variables contributed the most to the occurrence of a KSEV in a traffic accident, Table 9 did not include the variables in which the values of probabilities of SLIG were always higher than those of KSEV. In addition, it should be mentioned that none of the variables values in the original data set were found to be significantly affecting KSEV; however, and for comparison purposes only, values of variables in the original data set were shown for the same values which were found to be significant in the oversample data set.

Setting evidences for the values of variables used to build the BN indicated that VEH, PAT, DIR, ACT, LIG, SUP, and SPE were found to be significant. A detailed discussion of the most significant variables that were found to contribute to the occurrence of a killed or severe injury (KSEV) in a traffic accident is given below.

For both models in the oversample data set, accidents in which there was only one vehicle involved (VEH = 1) were found to be associated with KSEV occurrence. This result was consistent with the accident pattern variable (which also identified single vehicle as more significant in KSEV accidents) and the accident type variable in which the categories more associated with KSEV (collision with fall off vehicle, collision with fixed object and collision with animals) involved only one vehicle. Manner and Wünsch-Ziegler (2013) concluded that accidents due to collision with a roadside object tend to be more severe than others. In addition, Theofilatos et al. (2012) showed that collision with fixed objects was one of the most affecting factors of road accident severity in urban areas.

Undivided roads were associated with KSEV injuries; this result was consistent with the findings of Hosseinpour et al. (2014). Their results indicated that the presence of a median increased the probability of non-injury outcomes by 37%, while the probabilities of slight, severe, and fatal injuries decreased by 8.4%, 22.7%, and 5.9%, respectively.

The results also showed that when pavement surface value was others (conditions in which pavement was covered with mud, oil or sand) the severity of accident increased. Other researchers found that there was a relation between the surface status and resulting injury severity. Li et al. (2013) used skid resistance amongst other variables to investigate injury severity of accidents in Texas. They used a skid score, which is a value between one (least skid resistance) and ninety-nine (most skid resistance), that describes the overall skid resistance of the data collection section. They stated that the tests did not show many meaningful relationships for different groups of crashes; however, results obtained suggested that there was a significant correlation between skid scores and crash severity outcomes for highways with a dry surface condition, where pavements with lower skid resistance scores had more severe crashes than other skid scores.

According to Nevarez-Pagan (2008), driver crash involvements on locations with low skid resistance accounted for 28.56% of the severe injuries (total crashes) and 30.87% of the severe injuries (wet pavement crashes), however, the author recommended that the nature of these crashes (especially for wet pavement) should be further investigated. Herein, it should be emphasized that this specific condition was not prevalent (see Table 1), thus, it is suggested that the association between pavement surface condition and injury severity be further investigated.

In addition, it was found that accidents that occurred in a dark lighting condition were mostly associated with KSEV, this was also found by Yannis et al. (2013) where they concluded that the absence of street lighting during nighttime had the highest impact on the number of fatalities and severe injuries. Haleem and Gan (2011) analyzed traffic accidents on urban roadways; they concluded that the afternoon peak period showed the highest reduction in fatality/severity relative to night off-peak. On the other hand, compared to dark lighting conditions, daylight had the highest reduction. This result was consistent with results obtained by Ma et al. (2009), where they showed that night without lighting led to a higher risk of accident severity.

Speed values higher than 70 km/h were found to contribute to the occurrence of KSEV. This result was consistent with Manner and Wünsch-Ziegler (2013). Haleem and Gan (2011) also concluded that high speed limit sections experienced a significant increase in the occurrence of fatality/severity; while Kadilar (in press) found that the drivers travelling at speeds of more than 111 km/h have three times more risk than those travelling at speeds of less than 56 km/h.

4. Conclusions

The study presented in this paper investigated the possibility of using sampling techniques on imbalanced traffic accidents data sets prior to using different Bayes classifiers in order to develop models used to predict severity level of a traffic accident. All the data used in the study was obtained for urban and sub-urban roads in Jordan.

The overall conclusion that can be drawn from this study is that using balanced data sets improved the ability of Bayes classifiers to classify killed and severe injuries correctly while keeping the classification ability of the other class (i.e. slight injuries) in an acceptable level. On the other hand, Bayes models developed using an imbalanced data set were incapable of determining which values of variables were the most influencing to a killed or severe injury outcome in an accident.

Moreover, the results obtained using the original dataset were biased towards the majority class (i.e. slight injuries class) because of the dominating effect of the majority class, and so the classifier classified most instances as belonging to slight injuries class. However, the cost of misclassifying instances belonging to killed or severe injuries class far outweighs the cost of misclassifying the instances belonging to slight injury class. When using the classifier with balanced data set, the problem of costly misclassifications is much reduced to the extent that there was almost no bias towards any of the classes. More importantly, prediction of instances belonging to killed or severe injuries class was enhanced remarkably, especially when using the oversampling approach.

It was also found that models developed using BNs, specifically those developed using K2 and simulated annealing search algorithms were the best among other Bayes classifiers. Vehicle involved, accident type, accident pattern, direction, number of lanes, road section, grade, pavement type, surface condition, atmospheric condition, speed, traffic control and lighting conditions were all found to have a significant effect on severe outcomes in an accident.

In addition, inference was used to determine which specific values of variables contributed the most to the occurrence of a killed or severe injury in traffic accidents. The results of inference indicated that the following values of variables were found to be associated with higher probabilities of accidents with casualties including killed or severe injuries: single vehicles accidents, two way undivided roads, fall off vehicle, fixed object collisions, collision with animals, accidents occurring during dark lighting conditions, pavement surface covered with mud, sand or oil, and speed limit higher than 70 km/h. In general, these results were consistent with the literature (Haleem and Gan, 2011; Theofilatos et al., 2012; Manner and Wünsch-Ziegler, 2013; Hosseinpour et al., 2014).

Imbalanced data sets can be used to extract important knowledge when balanced, and since sampling techniques have proved their effectiveness in different research areas, this work indicated that they could also be applied in the domain of traffic accidents when used in conjunction with BNs. Their effectiveness has been found to be similar to other techniques used to model severity in traffic accidents. Compared with other well-known statistical methods, the main advantage of using sampling techniques and BNs seems to be their complex approach where misclassification cost is higher in one class than in the other and when there is no pre-defined underlying relationship between target variable and predictors needed (Chang and Wang, 2006; Chawla, 2005).

5. Recommendations

It should be emphasized that the effect of driver-related factors should be investigated to find out what implications they have on the severity of a traffic accident along with the factors studied herein. In addition, the authors recommend balancing traffic accidents data sets when learning from Bayes classifiers.

The results obtained herein might be used as guidance for PTD in order to develop countermeasures that aim to reduce the severity of traffic accidents in urban areas. They can also be used in safety awareness campaigns.

The authors of this research believe that collaboration between academic institutions and PTD is essential to help both apply and develop models that are used in the analysis of traffic accidents in order to determine the most influencing factors that contribute to the occurrence of a specific injury severity.

Acknowledgements

The authors are grateful to the Police Traffic Department in Jordan for providing the data necessary for this research. Griselda López wishes to express her acknowledgement to the regional ministry of Economy, Innovation and Science of the regional government of Andalusia (Spain) for their scholarship to train teachers and researchers in Deficit Areas, which has made this work possible. The authors appreciate the reviewers' comments and effort in order to improve the paper.

References

Abellán, J., De Oña, J., López, G., 2013. Analysis of traffic accident severity using decision rules via decision trees. Expert Syst. Appl. 40, 6047–6054.
Acid, S., de Campos, L.M., Fernández-Luna, J.M., Rodríguez, S., Salcedo, J.L., 2004. A comparison of learning algorithms for Bayesian networks: a case study based on data from an emergency medical service. Artif. Intell. Med. 30, 215–232.
Al-Ghamdi, A.S., 2002. Using logistic regression to estimate the influence of accident factors on accident severity. Accid. Anal. Prev. 34, 729–741.
Błaszczyński, J., Stefanowski, J., 2015. Neighbourhood sampling in bagging for imbalanced data. Neurocomputing 150 (Part B), 529–542.
Brown, I., Mues, C., 2012. An experimental comparison of classification algorithms for imbalanced credit scoring data sets. Expert Syst. Appl. 39, 3446–3453.
Chang, L.Y., Wang, H.W., 2006. Analysis of traffic injury severity: an application of non-parametric classification tree techniques. Accid. Anal. Prev. 38, 1019–1027.
Chawla, V.N., 2005. Data mining for imbalanced datasets: an overview. In: Maimon, O., Rokach, L. (Eds.), Data Mining and Knowledge Discovery Handbook. Springer US, pp. 853–867 (check citing a book chapter).
Chawla, N., Hall, L., Bowyer, K., Kegelmeyer, W., 2002. SMOTE: synthetic minority oversampling technique. J. Artif. Intell. Res. 16, 321–357.
Crone, S., Finlay, S., 2012. Instance sampling in credit scoring: an empirical study of sample size and balancing. Int. J. Forecast. 28, 224–238.
Daskalaski, S., Kopanas, I., Avouris, N., 2006. Evaluation of classifiers for an uneven class distribution problem. Appl. Artif. Intell. 20, 381–417.
De Oña, J., Mujalli, R.O., Calvo, F.J., 2011. Analysis of traffic accident injury severity on Spanish rural highways using Bayesian networks. Accid. Anal. Prev. 43, 402–411.
De Oña, J., López, G., Abellán, J., 2013. Extracting decision rules from police accident reports through decision trees. Accid. Anal. Prev. 50, 1151–1160.
Department of Statistics- DOS [online], 2013. Available from World Wide Web: http://web.dos.gov.jo/ (accessed March 2015).
Fawcett, T., Provost, F., 1997. Adaptive fraud detection. Data Min. Knowl. Discov. 1, 291–316.
Haleem, K., Gan, A., 2011. Identifying traditional and nontraditional predictors of crash injury severity on major urban roadways. Traffic Inj. Prev. 12, 223–234.
Helai, H., Chor, C.H., Haque, M.M., 2008. Severity of driver injury and vehicle damage in traffic crashes at intersections: a Bayesian hierarchical analysis. Accid. Anal. Prev. 40, 45–54.
Hosseinpour, M., Yahaya, A.S., Sadullah, A.F., 2014. Exploring the effects of roadway characteristics on the frequency and severity of head-on crashes: case studies from Malaysian Federal Roads. Accid. Anal. Prev. 62, 209–222.
Japkowicz, N., Myers, C., Gluck, M., 1995. A novelty detection approach to classification. In: Proceedings of the 14th international joint conference on Artificial intelligence, Montreal, Quebec, Canada, pp. 518–523.
Japkowicz, N., 2000. The class imbalance problem: significance and strategies. In: Proceedings of the 2000 International Conference on Artificial Intelligence (ICAI): Special Track on Inductive Learning, Las Vegas, Nevada.
Jiang, L., Zhang, H., 2006. Weightily averaged one-dependence estimators. In: Proceedings of the 9th Biennial Pacific Rim International Conference on Artificial Intelligence, PRICAI 2006, pp. 970–974.
Kadilar, G.O., 2015. Effect of driver, roadway, collision, and vehicle characteristics on crash severity: a conditional logistic regression approach. Int. J. Inj. Control Saf. Promot. (in press).

Khorashadi, A., Niemeier, D., Shankar, V., Mannering, F., 2005. Differences in rural and urban driver-injury severities in accidents involving large-trucks: an exploratory analysis. Accid. Anal. Prev. 37, 910–921.
Kockelman, K.M., Kweon, Y.J., 2002. Driver injury severity: an application of ordered probit models. Accid. Anal. Prev. 34, 313–321.
Kotsiantis, S., Kanellopoulos, D., Pintelas, P., 2006. Handling imbalanced datasets: a review. GESTS Int. Trans. Comput. Sci. Eng. 30, 25–36.
Kwon, O.H., Rhee, W., Yoon, Y., 2015. Application of classification algorithms for analysis of road safety risk factor dependencies. Accid. Anal. Prev. 75, 1–15.
Kubat, M., Holte, R.C., Matwin, S., 1998. Machine learning for the detection of oil spills in satellite radar images. Mach. Learn. 30, 195–215.
Li, H., Sun, J., 2012. Forecasting business failure: the use of nearest-neighbour support vectors and correcting imbalanced samples – evidence from the Chinese hotel industry. Tour. Manag. 33, 622–634.
Li, Y., Liu, C., Ding, L., 2013. Impact of pavement conditions on crash severity. Accid. Anal. Prev. 59, 399–406.
López, G., de Oña, J., Abellán, J., 2012a. Using decision trees to extract decision rules from police reports on road accidents. In: Proceedings of SIIV-5th International Congress – Sustainability of Road Infrastructures 2012, 53, pp. 106–114.
López, V., Fernández, A., Del Jesus, M.J., Herrera, F., 2012b. Cost sensitive and preprocessing for classification with imbalanced data-sets: similar behaviour and potential hybridizations. In: Proceedings of the 1st International Conference on Pattern Recognition Applications and Methods (ICPRAM 2012), Vilamoura, Portugal, pp. 98–107.
López, V., Fernandez, A., Garcia, S., Palade, V., Herrera, F., 2013. An insight into classification with imbalanced data: empirical results and current trends on using data intrinsic characteristics. Inf. Sci. 250, 113–141.
Ma, Z., Shao, C., Yue, H., Ma, S., 2009. Analysis of the logistic model for accident severity on urban road environment. In: Proceedings of IEEE Intelligent Vehicles Symposium (IV), pp. 983–987.
Madden, M.G., 2009. On the classification performance of TAN and general Bayesian networks. J. Knowl. Syst. 22, 489–495.
Manner, H., Wünsch-Ziegler, L., 2013. Analyzing the severity of accidents on the German Autobahn. Accid. Anal. Prev. 57, 40–48.
Milton, J.C., Shankar, V.N., Mannering, F.L., 2008. Highway accident severities and the mixed logit model: an exploratory empirical analysis. Accid. Anal. Prev. 40, 260–266.
Mittal, A., Kassim, A., Tan, T., 2007. Bayesian Network Technologies: Applications and Graphical Models. IGI Publishing, New York.
Montella, A., Aria, M., D'Ambrosio, A., Mauriello, F., 2012. Analysis of powered two-wheeler crashes in Italy by classification trees and rules discovery. Accid. Anal. Prev. 49, 58–72.
Mujalli, R.O., De Oña, J., 2011. A method for simplifying the analysis of traffic accidents injury severity on two-lane highways using Bayesian networks. J. Saf. Res. 42, 317–326.
Mujalli, R.O., De Oña, J., 2012. Injury severity models for motor vehicle accidents: a review. Proc. ICE – Transp. 166 (5), 255–270.
Neapolitan, R.E., 2009. Probabilistic Methods for Bioinformatics. Morgan Kaufmann Publishers, San Francisco, CA.
Nevarez-Pagan, A., Unpublished Master's thesis for master degree 2008. Severity of driver crash involvement on multilane high speed arterial corridors. University of Central Florida, Orlando, FL, USA.
Pahukula, J., Hernandez, S., Unnikrishnan, A., 2015. A time of day analysis of crashes involving large trucks in urban areas. Accid. Anal. Prev. 75, 155–163.
Pande, A., Abdel-Aty, M., 2009. Market basket analysis of crash data from large jurisdictions and its potential as a decision supporting tool. Saf. Sci. 47, 145–154.
Police Traffic Department - PTD [online], 2013. Available from World Wide Web: http://www.traffic.psd.gov.jo/ (accessed March 2015).
Poole, D., Mackworth, A., 2010. Artificial Intelligence: Foundations of Computational Agents. Cambridge University Press.
Qiu, C., Wang, C., Fang, B., Zuo, X., 2014. A multiobjective particle swarm optimization-based partial classification for accident severity analysis. Appl. Artif. Intell. 28, 555–576.
Savolainen, P., Mannering, F., Lord, D., Quddus, M., 2011. The statistical analysis of highway crash-injury severities: a review and assessment of methodological alternatives. Accid. Anal. Prev. 43, 1666–1676.
Simoncic, M., 2004. A Bayesian network model of two-car accidents. J. Transp. Stat. 7, 13–25.
Stefanowski, J., Wilk, S., 2008. Selective pre-processing of imbalanced data for improving classification performance. In: Proceedings of 10th International Conference DaWaK 2008, Lecture Notes in Computer Science, 5182, Springer, Berlin, Heidelberg, pp. 283–292.
Thammasiri, D., Delen, D., Meesad, P., Kasap, N., 2014. A critical assessment of imbalanced class distribution problem: the case of predicting freshmen student attrition. Expert Syst. Appl. 41, 321–330.
Theofilatos, A., Graham, D., Yannis, G., 2012. Factors affecting accident severity inside and outside urban areas in Greece. Traffic Inj. Prev. 13, 458–467.
Wang, H., Xu, Q., Zhou, L., 2015. Large unbalanced credit scoring using lasso-logistic regression ensemble. PLOS ONE 10 (2), 1–20.
Webb, G., Boughton, J., Wang, Z., 2005. Not so naive Bayes: aggregating one-dependence estimators. Mach. Learn. 58, 5–24.
Witten, I.H., Frank, E., 2005. Data Mining: Practical Machine Learning Tools and Techniques, 2nd ed. Morgan Kaufmann, San Francisco, CA.
World Health Organization- WHO [online], 2013. Global status report on road safety, Available from World Wide Web: http://www.who.int/mediacentre/factsheets/fs358/en/ (accessed February 2015).
Wu, J., Cai, Z., 2011. Learning averaged one-dependence estimators by attribute weighting. J. Inf. Comput. Sci. 8 (7), 1063–1073.
Yannis, G., Kondyli, A., Mitzalis, N., 2013. Effect of lighting on frequency and severity of road accidents. Proc. Inst. Civ. Eng.: Transp. 166, 271–281.
Zheng, F., Webb, G.I., 2006. Efficient lazy elimination for averaged one-dependence estimators. In: Proceedings of the 23rd international conference on Machine learning (ICML '06), New York, USA, pp. 1113–1120.