XIAOXIA XIONG, M.S.[1]
E-mail: x_xiong623@163.com
LONG CHEN, Ph.D.[1]
(Corresponding author)
E-mail: chenlong@ujs.edu.cn
JUN LIANG, Ph.D.[1]
E-mail: liangjun@ujs.edu.cn
[1] School of Automotive and Traffic Engineering,
Jiangsu University
301 Xuefu Road, Zhenjiang 212013, P.R. China

Safety and Security in Traffic
Preliminary Communication
Submitted: 28 Mar. 2017
Accepted: 27 Nov. 2017

ANALYSIS OF ROADWAY TRAFFIC ACCIDENTS BASED ON ROUGH SETS AND BAYESIAN NETWORKS

ABSTRACT

The paper integrates Rough Sets (RS) and Bayesian Networks (BN) for roadway traffic accident analysis. RS reduction of attributes is first employed to generate the key set of attributes affecting accident outcomes, which are then fed into a BN structure as nodes for BN construction and accident outcome classification. Such RS-based BN framework combines the advantages of RS in knowledge reduction capability and BN in describing interrelationships among different attributes. The framework is demonstrated using the 100-car naturalistic driving data from Virginia Tech Transportation Institute to predict accident type. Comparative evaluation with the baseline BNs shows the RS-based BNs generally have a higher prediction accuracy and lower network complexity while with comparable prediction coverage and receiver operating characteristic curve area, proving that the proposed RS-based BN overall outperforms the BNs with/without traditional feature selection approaches. The proposed RS-based BN indicates the most significant attributes that affect accident types include pre-crash manoeuvre, driver's attention from forward roadway to centre mirror, number of secondary tasks undertaken, traffic density, and relation to junction, most of which feature pre-crash driver states and driver behaviours that have not been extensively researched in literature, and could give further insight into the nature of traffic accidents.

KEY WORDS

roadway traffic accident; Rough Sets; Bayesian Networks; naturalistic driving; driver behaviour;

1. INTRODUCTION

Road traffic injuries and deaths have been a major public health issue globally. According to World Health Organization (WHO), approximately 1.25 million people die from roadway traffic accidents each year, while 20~50 million people suffer non-fatal injuries with many resulting in disabilities [1]. Accordingly, more efforts should be made to identify accident risk factors and reduce accidents. Naturalistic driving data (NDD), which are "collected from a number of equipped vehicles driven under naturalistic conditions over an extended period of time" [2], has shown great potential in the field and has attracted increasing attention over the last decade [3-9]. Different from empirical data collection by simulators or test tracks, NDD features natural/real driving behaviours of study participants (the instrumentation in the equipped vehicle is designed to be unobtrusive and no special instructions would be given to the participants), making the data more suitable to analyze the nature of traffic accidents. Besides, NDD covers a wide range of variables for safety-critical cases, such as pre-crash driver manoeuvres and driver's inattention and distraction just prior to the crash/near-crash, which are believed to be important factors contributing to traffic accidents [2,10-11] but rarely available in other resources. However, earlier works based on NDD have been focused on regression models including linear regression, Poisson regression, and logistic regression models [5-9], while such models predefine the underlying relationships between dependent and independent variables to be linear, which is a strict assumption that may be violated in application [12-13]. In addition, existing studies conducted on NDD have not emphasized the pre-selection of safety-critical features for model establishment, while the large number of pre-crash driver behaviour and environmental variables from NDD could pose a difficulty in developing an efficient and accurate model. Hence, Rough Sets (RS) and Bayesian Networks (BN), two kinds of popular data mining techniques, are introduced in the paper to address these deficiencies in literature.

Rough Sets (RS), proposed by Pawlak (1982) as an extension to set theory [14], derives decision or classification rules based on knowledge/attribute reduction and has proven to be useful in exploring data patterns or knowledge discovery [15]. It has the advantages of relaxing assumptions on the statistical nature of data (such as the independence of irrelevant alternatives assumption in multinomial logit model) and avoiding

structural constraints on the relationship between independent and dependent variables (like linear relations using regression). Previous works [16, 17] have explored Rough Sets (RS) in traffic accident analysis and achieve promising results. However, knowledge discovered by RS is expressed in the form of IF-THEN statements (decision rules) and could not describe possible interrelationships among different variables.

Bayesian Networks (BN) is also a data mining methodology that is being widely used for analyzing roadway traffic accidents [18-23]. Like RS, neither does BN impose any constraints on underlying relationships between the variables. Furthermore, BN could statically describe the interrelations between different variables in a tree-like structure, and make predictions based on these discovered relationships. However, BN usually requires a large data set and is complex in calculating especially when the number of variables increases [24]. As a result, variable or attribute selection is of great importance to BN construction. However, little attention has been paid to feature selection prior to BN construction in the existing roadway traffic accident applications. Filter and wrapper approaches have been typical feature selection for different domains. Wrapper approach, also called closed-loop method, employs a predictor performance as the criterion for feature selection and contains feedback from the model/algorithm that is to be used [25]. Thus, wrapper approach is required for BN analysis as a BN classifier has to be learned (which involves complicated structure learning and parameter learning that would be explained in detail in the following section) each time a subset of features is tested. Also, BN feature selection evaluation may also be complicated using wrapper approach due to the compound effect of different feature subsets and BN structure combination (a different BN structure should be learned for a different subset of features). Filter approach, also called open-loop method, on the other hand, relies on between-class separability and is implemented independent of the construction of the classifier, and thus is more appropriate for BN analysis in the domain [26]. However, typical filter approaches may have limitations in different aspects, like the biases arising from the number of categories one variable has for mutual information-based filter approach and due to linear assumptions embedded in the measurement for Pearson coefficient-based filter approach. As RS has shown effectiveness in attribute reduction while keeping the information system's classification capability unchanged in other areas [15], attribute reduction of RS is selected and explored as a variable pre-selection procedure for BN analysis of roadway traffic accidents in the paper.

As discussed above, although RS and BN are two popular techniques that have been applied to road traffic accident analysis, there are limitations of both methods, and applications combining the RS and BN in traffic accident analysis based on NDD have been rarely demonstrated. The paper proposes a combination of the two methods in analyzing traffic accidents using NDD. The proposed method is demonstrated using the 100-car naturalistic driving data from Virginia Tech Transportation Institute, which include many pre-crash driver state and behaviour data that have rarely been explored in literature. The proposed method is compared with baseline BNs with/without traditional filter approaches, and results prove the value in combining RS and BN in NDD-based roadway traffic accident analysis.

The rest of the paper is organized as follows. RS and BN theories are introduced in Section 2. Real NDD dataset adopted to demonstrate the proposed framework is described in Section 3. Results obtained are discussed in Section 4 and conclusions are drawn in Section 5.

2. METHODOLOGY

2.1 Baseline attribute reduction methodologies

As typical filter approaches utilized in feature selection for BN analysis, correlation analysis and mutual information analysis [23, 27] are employed as baseline feature selection methods for evaluating the performance of RS-based approach. Pearson's correlation coefficient is tested to check linear dependency of features on the decision attribute, while statistical $2N \cdot I(X,Y)$ using mutual information is formed and tested to measure their relationship from the perspective of information/entropy, where $I(X,Y)$ represents the mutual information of a condition attribute X and decision attribute Y. Statistically, $2N \cdot I(X,Y)$ asymptotically follows a $\chi^2_{(r_i-1)(r_0-1)}$ distribution, where N is the number of cases in the sample, r_i and r_0 are the numbers of values of the X and Y variable, respectively [27], $I(X,Y)$ could be calculated as follows:

$$I(X,Y) = \sum_{x,y} p(x,y) \log \frac{p(x,y)}{p(x)p(y)} \quad (1)$$

Only the variables that pass the correlation test and mutual information test at the 0.05 significance level are included in the baseline selective feature subsets for comparative evaluation.

2.2 RS attribute reduction

In RS theory, for a decision attribute Y (usually referred to as dependent variable in statistics) with n categories $\{Y_1, Y_2, ..., Y_i, ...Y_n\}$ (i.e., the universe $U = \bigcup_{i=1}^{n} Y_i$ classified into n non-overlapping classes), each of its category is described through lower approximation se $\underline{A}Y_i$ and upper approximation set $\overline{A}Y_i$ formed by elementary sets (as defined in *Equations 2* and *3*). An elementary set X is a set consisting of objects/cases that

are indiscernible by the specified set of condition attributes (usually referred to as independent variables in statistics), which represents the smallest partitions of cases given the specified condition attributes (i.e., cases from different elementary sets are discernible while those within the same elementary set are indiscernible).

$$\underline{A}Y_i = \bigcup X \quad \{X \in A^* \text{ and } X \subseteq Y_i\} \tag{2}$$

$$\overline{A}Y_i = \bigcup X \quad \{X \in A^* \text{ and } X \cap Y_i \neq \emptyset\} \tag{3}$$

where i=1,2,...n indicates different categories/classes of the decision attribute Y; A^* denotes the family of all elementary sets. *Equations 2* and *3* mean that the lower approximation set consists of all cases that are definitely within the category, while the upper approximation set also contains those not certainly belonging to the category.

Obviously, a category of decision attribute can be described differently by varying its lower and upper approximations through changing condition attributes. Two indicators are usually employed to evaluate the performance of the specified condition attributes in distinguishing cases, including accuracy of approximation and quality of approximation [16]. Specifically, accuracy of approximation reflects the performance of the set of condition attributes in discerning cases at categorical level, while quality of approximation reflects their overall discerning performance for the decision attribute. Detailed explanations of these two indicators are as follows:

1) Accuracy of approximation $\alpha(Y_i)$ is defined as the percentage of definable cases for i-th category of Y [16], expressed as:

$$\alpha(Y_i) = \frac{card(\underline{A}Y_i)}{card(\overline{A}Y_i)} \tag{4}$$

where *card* refers to cardinality. A larger $\alpha(Y_i)$ (ranging from 0 to 1) would indicate a higher discernibility of i-th category of Y given the set of condition attributes.

2) Quality of approximation $\gamma(Y)$ is defined as the percentage of definable cases for all categories of Y (i.e., the universe U) [16], expressed as:

$$\gamma(Y) = \frac{\sum_{i=1}^{n} card\,(\underline{A}Y_i)}{card\,(U)} \tag{5}$$

A $\gamma(Y)$ (ranging from 0 to 1) close to 1 implies that all categories of Y could be clearly identified by the specified set of condition attributes.

Attributes with poor performance in distinguishing cases based on such indicators are considered as redundant and should be excluded from the specified set of condition attributes, which process is referred to as attribute reduction in RS and yields a collection of reduct sets. A reduct set is regarded as the essential part of an information table which ensures the discernibility of all cases. More details in RS reduction of attributes could be found in [14].

In the paper, attributes existing in reduct sets are then employed as nodes in BN construction for classification, which would be described in detail in the following section.

2.3 BN analysis

BN is a Directed Acyclic Graph (DAG) model over a set of variables U={x_1,x_2,...,x_i,...x_n} which are represented by nodes of a network structure or graph. The interactions among the set of variables are represented by directed links (also referred to as arcs or edges) between the nodes, where node x_i is always pointed to from its parent nodes $pa(x_i)$. The direct influences implied by the directed links can be quantitatively described with a set of Conditional Probability Distributions (CPD) at each node:

$$CPD = \{Prob(x_i \mid pa(x_i)), x_i \in U\}, i = 1, 2, \ldots, n \tag{6}$$

It should be noted that as one can regard a link from node A to node B indicating that A causes B, the link may have different meanings not necessarily causal ones [21, 28], such as A being partial causation or predisposition of B, the two being functionally related, B being imperfect observation of A, or the two being statistically correlated [18]. Such meanings of a link are employed in the paper, that there exist some interrelations between nodes (variables) in a BN not limited to causal relationship.

Based on CPD, the joint probability distribution over the set of variables U represented by BN could be derived as follows:

$$P_{BN}(U) = \prod_{x_i \in U} Prob(x_i \mid pa(x_i)), i = 1, \ldots, n \tag{7}$$

A set of local independence assumptions are imposed for the factorization in *Equation 7*, which assert that in a BN each variable is independent of predecessors (from whom there exist directed path pointing into the node) given its parents [18]. Apparently, within BN, all variables are treated equally in a way that no dependent or independent variables need to be specified. When applied to the problem of classifying a variable of interest $x_c \in U$ (corresponding to decision attribute in RS) given data over the set of other variables in U (i.e., set of U excluding x_c; corresponding to condition attributes in RS), the probability of each category of x_c is calculated based on CPD and the class of x_c would be assigned to the category with highest probability.

A BN is typically constructed with the following three steps:

Step 1: Determine all relevant variables in constructing BN. RS reduction of attributes is employed in this step in the paper.

Step 2: Learn the structure of BN from data. There are basically two approaches in structure learning of BN, one is a searching and scoring-based method and the other a constraint-based method [18]. As previous studies [20-22, 29] indicate the hill-climbing search algorithm based on Minimum Description Length (MDL) score shows good classification accuracy while with relatively low network complexity, the algorithm was selected for building BN structure in the paper.

Step 3: Estimate parameters of BN from data. Given a BN structure, the conditional probability tables of the BN could be estimated from data based on maximum likelihood method or Bayesian approach [24]. To avoid possible overfitting problem arising from the large number of parameters, the Bayesian approach was adopted in the paper where the prior probability of each node of BN follows a Dirichlet distribution [24].

It should be noted that the last two steps (i.e., structure and parameter learning) interact and are carried out alternately to build a BN. Several indictors were selected for evaluating the built BN following previous research [21, 24, 30], including classification/prediction accuracy (ACC), absolute coverage percentage error (ACPE), area under receiver operating characteristic curve (AUC), and total number of BN arcs. Specifically, prediction accuracy and ACPE reflect the prediction performance at individual and aggregate level, respectively [30]; AUC reflects the overall performance of prediction by characterizing the trade-off between sensitivity and specificity [24]; the total number of BN arcs reflects the overall complexity of the built BN [21]. Detailed explanations of these indicators are as follows.

Prediction accuracy ACC_i is defined as the ratio of the number of correctly predicted instances classified as accident type i (N_{pa_i}) over the total number of observed instances in accident type i class (N_i) [30], expressed as

$$ACC_i = \frac{N_{pa\_i}}{N_i} \cdot 100\% \quad (8)$$

Absolute coverage percentage error ACPE measures the relative difference of the prediction coverage to the full (100%) coverage of accident type i based on observation, where prediction coverage is defined as the ratio of the number of predicted instances assigned to accident type i class (including both correct and incorrect prediction assignments) (N_{pc_i}) over the total number of observed instances in accident type i class (N_i) [30].

$$ACPE_i = \left| \frac{N_{pc\_i}}{N_i} \cdot 100\% - 100\% \right| \quad (9)$$

Receiver operating characteristic curve ROC is a term arising from signal detection [24]. ROC for accident type i plots the true positive (instances observed in accident type i class also predicted to be in accident type i class) rate vs. the false positive (instances not observed in accident type i class but predicted to be in accident type i class) rate, with the area under the curve (AUC) indicating the overall prediction performance for accident type i: a low AUC at 0.5 indicating a valueless prediction while its maximum at 1.0 indicating a perfect prediction.

Total number of BN arcs represents the complexity of a BN structure, as more links may entail exponentially larger conditional probability table and make parameter estimation difficult [21]. As such, among BNs with similar prediction performance, a BN with a relatively small number of arcs is usually preferred.

3. DATA DESCRIPTION AND PREPARATION

The dataset used in the paper is from 100-car naturalistic driving study conducted by Virginia Tech Transportation Institute (VTTI) in the Northern Virginia / Washington, D.C. area from 2004 to 2005 [31]. Each vehicle in the study was fitted with a suite of sensors and cameras to collect real-time vehicle movement signals, driver driving status, as well as traffic and environment feature information. The study finally obtained approximately 2,000,000 vehicle miles and 43,000 hours of driving data, including 68 crashes and 760 near-crashes (situations requiring a rapid, severe evasive manoeuvre to avoid a crash) data that are open to public. Type of traffic accident is selected to be the decision attribute in the paper for demonstration of the proposed RS-based BN framework, and data on crash and near-crash cases are combined as the dataset for accident type analysis, considering the relative small sample size of crash cases in the database and the similar nature in crashes and near crashes [32]. Removal of missing attribute cases (no analyzed data due to missing video records) and special driving scenario cases (including entering/leaving parking area, making U-turn, etc.) yields a total of 711 sample cases for study in the paper.

All attributes/variables included in the study and their explanations are presented in *Table 1*, where *a1-a21* are condition attributes (*a1-a2* are driver characteristics; *a3-a10* are driver behavioural factors; *a11-a21* are environmental attributes) and *d* is the decision attribute. Specifically, *a4-a8* are attributes characterizing driver's inattention and distraction behaviours just prior to (i.e., within 3 seconds) the crash/near-crash: *a4* indicates whether the driver's eye glance area was away from forward roadway but rather to centre mirror inside the vehicle; *a5* indicates whether the driver's eye glance area was away from forward roadway but rather to left/right mirror outside the vehicle; *a6* indicates whether the driver's eye glance area was away from forward roadway but rather to left/right window of the vehicle; *a7* and *a8* indicate the scope and depth of the secondary tasks (i.e., all those other than driving

Table 1 – Attributes and categories

Attribute		Category
Code	Description	
a1:AGE	Age	1.Young (18-35); 2.Middle-aged (36-55); 3.Old (above 55)
a2:GEN	Gender	1.Male; 2.Female
a3:MNV	Pre-crash manoeuvre	1.Going straight, constant speed in traffic lane (travelling in lane at a longitudinal acceleration generally less than + 0.25 g); 2.Going straight, accelerating in traffic lane (travelling in lane at a longitudinal acceleration generally greater than + 0.25 g); 3.Going straight, decelerating in traffic lane (travelling in lane at a longitudinal acceleration generally less than - 0.25 g); 4.Lane changing; 5. Starting/Stopping in traffic lane; 6.Turning left/right at intersection
a4:ICM	Inattention- centre mirror	1.No; 2.Yes
a5:ILM	Inattention-left/right mirror	1.No; 2.Yes
a6:ILW	Inattention-left/right window	1.No; 2.Yes
a7:NST	#Secondary tasks	1.No secondary task; 2.One secondary task; 3.Two secondary tasks; 4.Three secondary tasks
a8:HST	Highest secondary task rank	1.No secondary task; 2.Simple secondary task; 3.Moderate secondary task; 4. Complex secondary task
a9:HOW	Hands on wheel	1.None; 2.One (left or right) hand only; 3.Both hands; 4.Unknown
a10:SBU	Seatbelt use	1.None used; 2.Lap/shoulder belt; 3.Unknown
a11:TRF	Traffic flow	1.Divided (median strip or barrier); 2.Not divided; 3.One-way traffic
a12:NTL	# Travel lanes	1.One lane; 2.Two lanes; 3.Three lanes; 4.Four lanes; 5. >= 5 lanes
a13:TRD	Traffic density	1.Level Of Service (LOS)-A; 2.LOS-B; 3.LOS-C; 4.LOS-D; 5.LOS-E & F
a14:TRC	Traffic control	1.No traffic control; 2.Traffic signal; 3.Stop sign; 4.Yield sign; 5.Traffic lanes marked (markings on the road that contain information or warnings applicable to the driving task); 6.Other (One-way road or street, officer or watchman, toll booths, etc.)
a15:RTJ	Relation to junction	1.Non-junction; 2.Intersection; 3.Intersection-related; 4.Interchange; 5. Entrance/exit ramp; 6.Other (Driveway, alley access, etc.)
a16:CUR	Curve	1.Roadway alignment is straight; 2.Roadway alignment is curved
a17:GRD	Grade	1.Flat (no hills or grade); 2.Grade (vehicle is going up or down a grade)
a18:LDU	Land use	1.Business/industrial; 2.Residential; 3.Open country; 4.Interstate; 5.Other (construction zone, school, church, etc.)
a19:LIG	Lighting	1.Dawn; 2.Daylight; 3.Dusk; 4.Darkness, lighted; 5.Darkness, not lighted
a20:WTH	Weather	1.Clear; 2.Cloudy; 3.Other (raining, snowing, mist)
a21:SUR	Surface	1.Dry; 2.Wet/Icy
d:ACT	Accident Type	1. Rear-end-striking (364); 2.Rear-end-struck (69); 3.Same direction sideswipe (111); 4.Opposite direction head-on/sideswipe (22); 5.Road departure (left or right) (55); 6.Intersection conflict/collision (66); 7.Forward conflict/collision with objects other than running vehicles (including parked vehicles, stationary objects, pedestrians, and animals) in roadway (24)

task) the driver was undertaking, where the rank of secondary tasks (i.e., simple, moderate, and complex) here follows the standards defined in [33].

4. RESULTS AND DISCUSSION

4.1 RS reducts

For comparative study, correlation test and mutual information test are implemented in Matlab to obtain baseline feature subsets, with the test results presented in *Table 2*. RS reduction of attributes are realized using genetic algorithm package within Rosetta software [34]. Finally, a total of 17 reducts were obtained as presented in *Table 3*, with a range of 15~16 attributes in each reduct set.

From *Table 3*, it could be noted that RS reducts include several variables that seem linearly independent of or irrelevant with accident type (ACT) from correlation analysis and mutual information analysis, such as driver's age (AGE), grade of roadway (GRD), and weather condition (WTH), indicating these variables may have indirect effects on accident type that are not reflected in correlation analysis nor mutual information

Table 2 – Correlation analysis and mutual information analysis results for accident type (ACT)

	Correlation Analysis		Mutual Information Analysis		
	Corr. Coef.	Corr_Sig.	$I(X_i,C)$	$2 \cdot N \cdot I$	I_Sig.
AGE	0.0039	0.9170	0.0123	17.512	0.1313
GEN	-0.0451	0.2300	0.0040	5.724	0.4547
MNV	0.0874	0.0198*	0.1325	188.345	0.0000†
ICM	-0.0500	0.1827	0.0341	48.420	0.0000†
ILM	-0.0402	0.2846	0.0390	55.445	0.0000†
ILW	-0.0664	0.0771	0.0134	19.054	0.0041†
NST	-0.0606	0.1065	0.0457	64.923	0.0000†
HST	-0.0926	0.0135*	0.0467	66.430	0.0000†
HOW	0.0917	0.0144*	0.0585	83.253	0.0000†
SBU	0.0185	0.6223	0.0148	21.091	0.0491†
TRF	0.2233	0.0000*	0.1020	145.070	0.0000†
NTL	-0.1663	0.0000*	0.1537	218.624	0.0000†
TRD	-0.3737	0.0000*	0.2102	298.907	0.0000†
TRC	0.1157	0.0020*	0.1004	142.785	0.0000†
RTJ	0.1135	0.0024*	0.2492	354.316	0.0000†
CUR	0.1361	0.0003*	0.0367	52.215	0.0000†
GRD	0.0486	0.1960	0.0062	8.850	0.1822
LDU	-0.0887	0.0180*	0.1594	226.661	0.0000†
LIG	0.0769	0.0403*	0.0565	80.298	0.0000†
WTH	0.0055	0.8846	0.0090	12.764	0.3865
SUR	0.0458	0.2223	0.0090	12.751	0.0472†

*Note: "Corr. Coef." represents Pearson correlation coefficients estimated between each attribute and ACT from the sample. "Corr_Sig." represents the significance level from the correlation test. "$2 \cdot N \cdot I$" represents the statistic $2N \cdot I(X,Y)$ which asymptotically follows a $\chi^2_{(r_i-1)(r_0-1)}$ distribution. "I_Sig." represents the significance level from the mutual information test. * indicates variables statistically significant at the 0.05 level in correlation test. † indicates variables statistically significant at the 0.05 level in mutual information test.*

Table 3 – Reduct sets based on rough set theory

Set No.	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17
AGE																	
GEN	×	×	×		×	×											
MNV*†																	
ICM†																	
ILM†	×	×	×	×	×	×	×	×	×	×	×	×	×	×	×	×	×
ILW†	×		×	×	×	×	×	×	×	×	×	×	×	×	×	×	×
NST†		×		×	×				×	×		×					
HST*†																	
HOW*†															×		
SBU†																	
TRF*†	×							×				×		×			
NTL*†		×		×		×	×										
TRD*†																	
TRC*†	×		×				×			×	×						×
RTJ*†																	
CUR*†		×	×	×	×	×	×		×				×		×	×	×
GRD																	
LDU*†																	
LIG*†											×		×	×		×	
WTH								×	×				×				×
SUR†	×	×	×	×	×	×	×	×		×	×	×		×	×	×	
#of Attributes	15	15	15	15	15	15	15	16	16	16	16	16	16	16	16	16	16

*Note: × indicates variables excluded from RS reduct sets. * indicates variables statistically significant at the 0.05 level in correlation test. † indicates variables statistically significant at the 0.05 level in mutual information test.*

analysis but could be beneficial to classification. Most inattention and distraction-related attributes are found to contribute to accident type based on mutual information and RS analysis, except for ILM and ILW (attention is paid to left/right mirrors and windows instead of forward roadway) from RS. It may be that eye stays at left/right areas of the vehicle and are dispensable in preserving the discernibility relation among cases within the same accident type and their removal would not worsen the classification given the dataset, while the inclusion of such variables may increase the complexity of a BN structure.

Different feature subsets are then fed into the BN structure as nodes for BN construction and classification. Obviously, such reduction in attributes would help reduce difficulty in BN analysis thanks to fewer nodes in a BN network. Their differences in prediction performances would be further explored and discussed in the following section.

4.2 BN based on RS reducts

4.2.1 Validation

Weka software [35] is used for BN structure and parameter learning. A 10-fold procedure [24] is adopted to avoid overfitting, where the whole dataset is divided into 10 exclusive folds and 10 rounds of training and testing procedures (for each round one of the 10 folds is held back for testing only) are performed for model learning and evaluation. To evaluate the performance of the proposed RS-based BN framework, a BN without RS reduction of attributes, a BN with mutual information-based filter and a BN with Pearson's correlation coefficient-based filter are selected as baselines for comparison with the BNs based on 17 reducts derived from RS (as presented in *Table 3*).

Table 4 summarizes and compares the values of performance indicators including accuracy, ACPE, and AUC for the baseline BNs and RS-based BNs (the indicator values for each RS subset are shown in *Figure 1*). The RS+BN framework outperforms all other baseline BNs in terms of prediction accuracy, while having a comparable prediction coverage (indicated by ACPE) with the baseline BN without RS attribute reduction and a similar AUC performance with the BNs with feature selection based on mutual information and

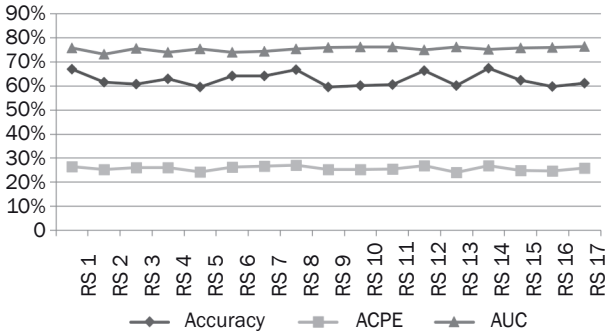

Figure 1 – Accuracy, ACPE, and AUC for RS-based BNs

Pearson correlation coefficient analysis. The network complexity (indicated by #BN arcs) of the RS-based BN is reduced by approximately 1/3 compared with the full attribute space (i.e., BN1).

Table 5 gives further comparison in performances of the baseline BN versus the RS-based BN with the highest prediction accuracy (that is BN3 in *Table 4* - a BN with mutual information-based filter versus the NO.1 reduct set in *Table 3*, referred to as RS-BN_{opt} further in the text) among different accident types. Results show that neither of them gives a perfect prediction accuracy rate for each accident type, especially for those without sufficient observations (i.e., accident types other than rear-end), and this problem has also been reported in other research works [16, 30]. The RS-BN_{opt} shows a better performance on prediction accuracy over the baseline BN3 in predicting each accident type. The overall prediction accuracy for all accident types is evaluated by a weighted average based on sample size (with the number of observations for each type being its weight) to counteract the negative impact of small sample size on prediction accuracy. The results show that the proposed RS-BN_{opt} has an 11% higher weighted average of accuracy (66.9% vs. 60.4%) compared to the baseline BN3 and thus also has a better performance on overall prediction accuracy. The same weighted average is calculated for ACPE and AUC, and results indicate that RS-BN_{opt} also performs slightly better in terms of AUC and prediction coverage (with a smaller ACPE). Finally, the RS-BN_{opt} has a slightly larger number of BN arcs, meaning the network generated from the proposed framework is slightly more complex than the one from the mutual information-based feature selection approach.

Table 4 – Performance indicators for different BNs

Performance Indicators	BN1 (Full set)	BN2 (Corr. Coef.)	BN3 (Mutual Info.)	RS-based BNs Averge±s.d.[1]
Accuracy [%]	60.3	59.6	60.4	62.6±2.8
ACPE [%]	25.9	28.1	26.6	25.7±0.9
AUC	0.745	0.760	0.752	0.752±0.009
#BN arcs	33	12	18	21±1

[1]*s.d.: standard deviation*

Table 5 – Performance indicators for baseline BN3 and RS-based BN (1)

Accident Type	BN3 (Mutual Info.)					RS-BN$_{opt}$			
	Accuracy [%]	ACPE [%][1]	AUC	#BN arcs		Accuracy [%]	ACPE [%]	AUC	#BN arcs
Rear-end-striking	76.4	|+17.9|	0.725			82.1	|+20.8|	0.717	
Rear-end-struck	66.7	|-66.7|	0.656			71.0	|-52.5|	0.681	
Same dir. sideswipe	50.5	|-22.9|	0.763			57.7	|-29.6|	0.758	
Head-on/ opposite dir. sideswipe	36.4	|-63.6|	0.715	18		40.9	|-57.3|	0.840	20
Road departure	23.6	|+23.6|	0.830			29.1	|+9.1|	0.891	
Intersection	30.9	|+24.8|	0.915			39.4	|+18.2|	0.917	
Other	33.3	|-37.5|	0.791			54.2	|-55.0|	0.772	
Weighted Avg.	60.4	26.6	0.752			66.9	26.4	0.758	

[1] *|a|: Absolute value of a. ACPE for each accident type is presented along with negative/positive signs before taking absolute values.*

From *Figure 1*, *Table 4*, and *Table 5*, the RS-based BN framework overall shows sizable improvement in prediction accuracy with relatively lower network complexity, while having comparable prediction coverage and AUC performance with the baseline BNs with/ without traditional feature selection (including mutual information and Pearson correlation coefficient-based approaches) given the dataset. Thus, the proposed RS-based BN is recommended to further analyze the interrelationship of attributes for traffic accidents.

4.2.2 Significance of attributes

The built RS-BN$_{opt}$ which has the best performance in prediction accuracy is chosen to explore the significance of different attributes in accident type classification. The learned structure of RS-BN$_{opt}$ is presented in *Figure 2*.

Figure 2 shows that among all the selected attributes from RS, seatbelt use, highest secondary task rank, and land use are the ones that are not directly related to accident type but with indirect influences (having links connected to attributes that directly connects with accident type). Specifically, seat belt use directly relates to age, which suggests there may be certain age groups that have tendencies in using/ not using seat belt; seat belt use also directly relates to hands on wheel indicator. This suggests that the seat belt usage also correlates with whether the driving is performed using both hands for driving, and this may reflect the driver's habit features like how cautious the driver is when driving, which would ultimately influence the outcome of an accident. The highest secondary task rank directly relates to driver's attention to centre mirror instead of forward roadway and the number of secondary tasks undertaken, which would directly affect the type of accident. Land use is correlated with relation to junction and the number of travel lanes, which may have a direct influence on traffic environment and consequently affect accident outcome. Hence, such BN structure could capture interesting interrelationships among different attributes that is not available in the RS decision rule framework.

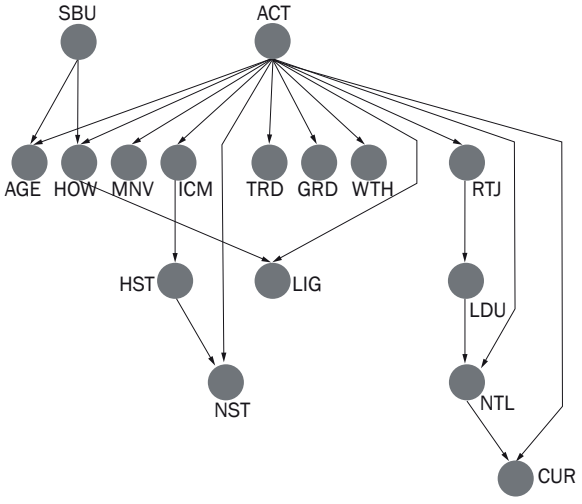

Figure 2 – RS-based BN Structure

Setting the evidence procedure [20, 21] is utilized to assist the identification of attributes and values that contribute most to the occurrence of each accident type, with the results presented in *Table 6*. For setting evidence procedure, each variable (discrete) of the built BN is set to one of its categorical value at a time, and the associated posterior probabilities of each accident type are calculated based on the Bayes rule (*Equation 10*) and CPD of the built BN.

$$P(d=i \mid a_k=j)=\frac{P(a_k=j \mid d=i)\cdot P(d=i)}{P(a_k=j)}= \\ =\frac{P(a_k=j \mid d=i)\cdot P(d=i)}{\sum_{i=1}^{7} P(a_k=j \mid d=i)\cdot P(d=i)} \quad (10)$$

where i refers to the type index of accidents, j refers to the category index of condition attribute a_k, $P(a_k=j \mid d=i)$ are obtained from the CPD of the built BN.

Table 6 – Inference results for attributes directly related to accident type

		d=1[1]	*d*=2	*d*=3	*d*=4	*d*=5	*d*=6	*d*=7
a3:MNV	*a3*=4	0.3241	0.1568	0.4582	0.0175	0.0187	0.0188	0.0059
	a3=5	0.2714	0.1424	0.1968	0.0604	0.0128	0.2797	0.0365
	a3=6	0.1311	0.0142	0.2438	0.0924	0.2666	0.2120	0.0399
a4:ICM	*a4*=2	0.3761	0.1738	0.3772	0.0090	0.0274	0.0274	0.0090
a7:NST	*a7*=4	0.2464	0.1134	0.3160	0.0516	0.0517	0.1743	0.0466
a13:TRD	*a13*=1	0.2751	0.0771	0.0606	0.0513	0.2776	0.1740	0.0842
a15:RTJ	*a15*=2	0.3319	0.0815	0.0826	0.0628	0.0316	0.3929	0.0167
	a15=5	0.3756	0.0814	0.4186	0.0084	0.0986	0.0090	0.0085

[1] *Meanings of codes and their index values are explained in Table 1. The highest probability value for each evidence set is in italics.*

Due to the limitation of paper space, identified significant attributes are limited to those directly related to accident type (i.e., there is a direct link between such attributes and the accident type). Also, as the type of rear-end-striking accident is usually positively predicted given the dataset (with the highest positive coverage percentage error at +20.8%, as shown in *Table 5* for RS-based BN), *Table 6* does not include the attributes and values that always yield higher probability values for the rear-end-striking type than for other accident types.

Table 6 shows that pre-crash manoeuvre, driver's attention from forward roadway to centre mirror, number of secondary tasks undertaken, traffic density, and relation to junction are significant attributes that directly relate to accident types other than rear-end-striking. Results indicate that manoeuvres including lane changing, starting/stopping in traffic lane, and turning (both left and right) at intersections are more likely to be involved in sideswipe (same direction), intersection collision, and road departure accidents, respectively. Driver's attention to centre mirror instead of forward roadway is found to be associated with sideswipe (same direction) accident, probably because a relatively long eye stay on the centre mirror would make the driver less capable of noticing vehicles on either side of the road. Having more than two secondary tasks is also a significant factor in sideswipe (same direction) accident, suggesting that sideswipe (same direction) accidents are usually related to inattention and distraction. Low traffic density is found to be more likely to contribute to road departure accidents, which may be due to high travelling speed that frequently occurs in low density traffic environment. Entry/exit ramp infrastructures are found to be more significant in sideswipe (same direction) accident, which is consistent with what one would expect as more demanding lateral driving manoeuvres are required at such facilities. The results indicate the importance of interventions targeting driver behaviours and suggest the potential value of driver monitoring system (such as online detection of driving manoeuvres and driver's inattention and distraction) in predicting/preventing different traffic accident types.

5. CONCLUSION

In this paper, a new framework integrating Rough Set (RS) and Bayesian Networks (BN) is proposed to analyze information from traffic accident database. In the proposed framework, RS reduction of attributes is first employed to generate the key set of attributes affecting accident outcomes, which are then fed into a BN structure as nodes for BN construction and accident outcome classification. Such framework combines the advantages of RS in knowledge reduction capability and BN in describing interrelationships among different attributes. The framework is demonstrated using the 100-car naturalistic driving data from Virginia Tech Transportation Institute to predict the accident type. Comparative evaluation with the baseline BNs shows that the RS-based BNs generally have a higher prediction accuracy and lower network complexity, while with comparable prediction coverage and ROC curve area, it proves that the proposed RS-based BN overall outperforms the BNs with/without traditional feature selection approaches. Also, the most significant attributes identified that affect accident types include pre-crash manoeuvre, driver's attention from forward roadway to centre mirror, a number of secondary tasks undertaken, traffic density, and relation to junction. Most of these attributes feature pre-crash driver states and driver behaviours that have rarely been studied in the existing literature based on BN [18-23], which could give further insight into the nature of traffic accidents.

The paper is a new attempt to apply the RS-based BN as a complementary tool for roadway traffic accident analysis based on NDD. The contribution of the proposed method to the literature is twofold. Firstly, the study here adds to the currently limited body of research regarding safety-critical feature selection method using NDD, which features a wide variety of driver state/behaviour and environmental variables

that may pose challenges for establishing an efficient and accurate model. Secondly, compared with the many previous NDD-based studies on driver state/behaviour-driving safety relationship that applied regression models allowing estimating of their direct effects only, the BN explored in the paper allows extracting direct and indirect relationships from the selected variables and could better reflect the complexity of the driver-vehicle-environment system.

Except for accident type demonstrated in the paper, other accident outcomes such as accident occurrence and accident severity can be examined using the same approach. More advanced models such as variable precision rough set (VPRS) and other BN structure searching and scoring algorithms should be considered in the future to improve the prediction performance of the framework. More sample cases with more details in pre-crash driving behaviour from naturalistic driving should also be collected in the future to further explore the crash-avoidance strategies.

ACKNOWLEDGEMENTS

This work has been jointly supported by the National Natural Science Foundation of China (NSFC) under grant U1564201 and 61773184, China Postdoctoral Science Foundation under 2016M600375, and Six talent peaks project in Jiangsu Province under 2015-DZXX-048.

熊晓夏[1]
E-mail: x_xiong623@163.com
陈龙[1]
E-mail: chenlong@ujs.edu.cn
梁军[1]
E-mail: liangjun@ujs.edu.cn
[1] 江苏大学汽车与交通工程学院
江苏省镇江市学府路301 号212013
中国

基于粗糙集和贝叶斯网络的道路交通事故分析

摘要

本文将粗糙集 *(RS)* 和贝叶斯网络 *(BN)* 结合起来进行道路交通事故分析。首先利用*RS*属性约简生成影响事故结果的关键属性集，然后将其作为*BN*结构的节点进行事故结果分类。这种基于*RS*和*BN*的综合分析方法结合了*RS*在知识约简能力和*BN*描述不同属性之间相互关系方面的优势。利用弗吉尼亚理工大学交通研究所的*100*车驾驶数据对事故类型进行了预测，与基础模型比较结果表明，基于*RS*的*BN*通常具有更高的预测精度和更低的网络复杂性且具有可比的预测范围和*ROC*曲线面积，证明了基于*RS*的*BN*整体优于采用传统特征选择方法或不采用特征选择方法的*BN*。模型结果表明影响事故类型最重要的属性包括驾驶员碰撞前操作，驾驶员注意力由前方道路转移至后视镜，非驾驶任务数目，交通流密度，与交叉口位置关系，其中大部分属性刻画了在文献中还没有得到广泛研究的碰撞前驾驶员状态和行为，有助于进一步探索交通事故的发生机理。

关键词

道路交通事故；粗糙集；贝叶斯网络；自然驾驶；驾驶员行为

REFERENCES

[1] World Health Organization. *Road Traffic Injuries.* World Health Organization; 2016. Available from: http://www.who.int/mediacentre/factsheets/fs358/en/

[2] Klauer S, et al. *The Impact of Driver Inattention on Near-Crash/Crash Risk: An Analysis Using the 100-Car Naturalistic Driving Study Data*. Washington D.C.: U.S. Department of Transportation; 2006.

[3] Piccinini GB, Engström J, Bärgman J, et al. Factors contributing to commercial vehicle rear-end conflicts in China: A study using on-board event data recorders. *Journal of Safety Research*. 2017;62: 143-153.

[4] Wang W, Liu C, Zhao D. How Much Data is Enough? A Statistical Approach with Case Study on Longitudinal Driving Behavior. *IEEE Transactions on Intelligent Vehicles*. 2017;2(2): 85-98.

[5] Precht L, Keinath A, Krems JF. Identifying the main factors contributing to driving errors and traffic violations – Results from naturalistic driving data. *Transportation Research Part F: Traffic Psychology & Behaviour*. 2017;49: 49-92.

[6] Xiong H, Bao S, Sayer J, et al. Examination of drivers' cell phone use behavior at intersections by using naturalistic driving data. *Journal of Safety Research*. 2015;54: 89.e29-93.

[7] Precht L, Keinath A, Krems J F. Effects of driving anger on driver behavior – Results from naturalistic driving data. *Transportation Research Part F: Psychology & Behaviour*. 2017;45: 75-92.

[8] Wu KF, Aguerovalverde J, Jovanis PP. Using naturalistic driving data to explore the association between traffic safety-related events and crash risk at driver level. *Accident Analysis & Prevention.* 2014;72: 210-218.

[9] Wang Y, Zhang W. Analysis of Roadway and Environmental Factors Affecting Traffic Crash Severities. *Transportation Research Procedia*. 2017;25: 2124-2130.

[10] Zhang G, Yau KK, Zhang X, et al. Traffic accidents involving fatigue driving and their extent of casualties. *Accident Analysis & Prevention.* 2016;87: 34-42.

[11] Talbot R, Fagerlind H, Morris A. Exploring inattention and distraction in the SafetyNet Accident Causation Database. *Accident Analysis & Prevention.* 2013;60(45): 445-455.

[12] Chang L, Wang H. Analysis of traffic injury severity: An application of non-parametric classification tree techniques. *Accident Analysis & Prevention.* 2006;38(5): 1019-1027.

[13] Zong F, Xu H, Zhang H. Prediction for Traffic Accident Severity: Comparing the Bayesian Network and Regression Models. *Mathematical Problems in Engineering*. 2013;2013(2-3): 206-226.

[14] Pawlak Z. Rough sets. *International Journal of Parallel Programming*. 1982;11(5): 341-356.

[15] Bello R, Falcon R. Rough Sets in Machine Learning: A Review. *Thriving Rough Sets, Studies in Computational Intelligence*. 2017;708: 87-118.

[16] Wong J, Chung Y. Rough set approach for accident

chains exploration. *Accident Analysis & Prevention*. 2007;39(3): 629-637.

[17] Peng L, Chao-Zhong W, Huang Z. Situation Assessment of Vehicle Collision Risk Based on Variable Precision Rough Set. *Journal of Transportation Systems Engineering & Information Technology*. 2013;13(5): 120-126.

[18] Simoncic M. A Bayesian network model of two-car accidents. *Journal of Transportation & Statistics*. 2004;7(1): 3594-3597.

[19] Karimnezhad A, Moradi F. Road accident data analysis using Bayesian networks. *Transportation Letters: The International Journal of Transportation Research*. 2015.

[20] Oña J, López G, Mujalli R, et al. Analysis of traffic accidents on rural highways using Latent Class Clustering and Bayesian Networks. *Accident Analysis & Prevention.* 2013;51C(2): 1-10.

[21] Oña J, Mujalli R, Calvo F. Analysis of traffic accident injury severity on Spanish rural highways using Bayesian networks. *Accident Analysis & Prevention.* 2011;43(1): 402-11.

[22] Mujalli R, De O. A method for simplifying the analysis of traffic accidents injury severity on two-lane highways using Bayesian networks. *Journal of Safety Research*. 2011;42(5): 317-326.

[23] Guo X, Zhang H, Fang Z. Bayesian network modeling for causation analysis of traffic accident. *Journal of Jilin University (Engineering and Technology Edition)*. 2011;41: 89-94.

[24] Bishop C. *Pattern Recognition and Machine Learning*. New York: Springer-Verlag; 2006.

[25] Cios K, Pedrycz W, Świniarski R. *Data Mining Methods in Knowledge Discovery*. London: Kluwer Academic Publishers; 1998.

[26] John G, Kohavi R, Pfleger K. Irrelevant features and the subset selection problem. In: Cohen WW, Hirsh H. (eds.) *Machine Learning: Proceedings of the 11th International Conference of Machine Learning, July 10–13 1994, Rutgers University, New Brunswick, NJ*. San Francisco CA: Morgan Kaufmann Publishers; 1994. p. 121-129.

[27] Blanco R, Inza I, Merino M, et al. Feature selection in Bayesian classifiers for the prognosis of survival of cirrhotic patients treated with TIPS. *Journal of Biomedical Informatics.* 2005;38(5): 376-388.

[28] Neapolitan R. *Probabilistic Methods for Bioinformatics*. San Francisco: Morgan Kaufmann Publishers; 2009.

[29] Madden M. On the classification performance of TAN and general Bayesian networks. *Knowledge-Based Systems*. 2009;22(7): 489-495.

[30] Cheng L, Chen X, Wei M, et al. Modeling mode choice behavior incorporating household and individual sociodemographics and travel attributes based on rough sets theory. *Computational Intelligence & Neuroscience*. 2014;2014: 1-9.

[31] Virginia Tech Transportation Institute. *VTTI Data Warehouse*. Virginia Tech Transportation Institute; 2016. Available from: http://forums.vtti.vt.edu/index.php?/files/category/3-100-car-data/

[32] Guo F. Near-Crashes as Crash Surrogate for Naturalistic Driving Studies. Transportation *Research Record Journal of the Transportation Research Board*. 2010;2147(-1): 66-74.

[33] Dingus T, Hulse M, Antin J, et al. Attentional demand requirements of an automobile moving-map navigation system. *Transportation Research Part A: Policy and Practice*. 1989;23(4): 301-315.

[34] Komorowski E. ROSETTA- A Rough Set Toolkit for Analysis of Data. *Proceedings of 3rd International Joint Conference on Information Sciences*; 1997. p. 303-407.

[35] Witten I, Frank E. *Data Mining: Practical Machine Learning Tools and Techniques*. 2nd ed. San Francisco: Morgan Kaufmann; 2005.