A road accident pattern miner (RAP miner)

S. M. N. Arosha Senanayake\textsuperscript{a} and Sisir Joshi\textsuperscript{b}

\textsuperscript{a}Institute of Applied Data Analytics, Universiti Brunei Darussalam, Gadong, Brunei; \textsuperscript{b}Information Technology, Charles Sturt University, Sydney, Australia

\textbf{ABSTRACT}

Domain-specific data service models can retrieve critical features from frequently occurring road accident patterns (RAPs). The aim of this research is to propose scan efficient association rules’ mining-based pattern analysis which provides more accurate RAP prediction in frequent accident locations with the fastest matching pattern search from a RAP database (RAP DB). Association rules’ mining technique derives a correlation between frequent RAP and association among various attributes of a road accident. While the clustering technique discriminates different RAPs, Naïve Bayes Classification classifies and then predicts the severity of accident using Fuzzy Inference Engine (FIE) interfaced with RAP Case Library (RAP CL) using hybrid intelligence. The results of the proposed road accident data service model prove a significant increase in the accuracy of accident prediction compared to the reported results. A novel hybrid learning algorithm, interfaced with Scan Efficient Apriori (SEA) algorithm implemented, leads the fast RAP search from the first scan through RAP CL and retain new RAP in the RAP CL using case-based reasoning (CBR) during subsequent scanning. Thus, the RAP miner built proves road accident prediction using SEA, FIE and CBR with the highest accuracy and fast RAP set processing.

\textbf{ARTICLE HISTORY}

Received 14 April 2021
Accepted 12 July 2021

\textbf{KEYWORDS}

Road accident patterns; association rules mining; frequent pattern; case-based reasoning; pattern-based classification

\section*{Introduction}

The main aim of this research is to introduce a novel Road Accident Pattern (RAP) miner model to retrieve matching RAP set in real time, online, and offline based on a knowledge base already established using historical pattern set available. RAP miner uses a domain-specific data model which extracts matching RAP based on the pattern fusion. The intelligent software built and embedded in RAP miner can easily adapt, revise and reuse for new occurrences of road accidents anywhere at any time using the case library established in this work, which allows setting up (if new), enhancing (if already existing), and revising (if required the repair) using case-based reasoning (CBR).

Thus, research problem addressed in this article is to find out appropriate pattern fusion techniques to revise/repair, reuse and retain matching RAPs in a case library.
already established using pre-accident (no accident), during accident (if any), and post-accident pattern sets by extracting a relevant pattern set from the existing RAP database. This will allow road authorities to apply rehabilitation techniques for roads/drivers and performance optimization of drivers involved in accidents. Thus, injury prevention of road users is expected in future occurrences.

The significant contribution of this article is to design and develop a novel pattern miner for RAPs, by proving the accuracy and precision in the prediction of road accidents using a well-established road accident database which can be applied by the road accident prevention authority of any country/city/urban areas.

Hence, the main goal of this work is to develop a RAP miner (RAP miner) to increase the accuracy and precision in the prediction of road accidents and decrease the processing time leading to prevention measures and hence, the least severity of injuries. The use of single technique has failed to predict with high accuracy and precision thus, hybrid learning algorithms built in this research, based on matching pattern mining, facilitate the decision-making of road accidents’ prediction with the least severity of injuries. These outcomes allowed increasing prevention measures by relevant authorities before, during or after a road accident.

**Related work**

Road accident occurs anywhere at any time. As a nature of human being, public safety and avoid accidents are primary in daily lifestyle. Big data analytics of road accident data deriving safe-driven features is crucial. Data mining, using domain-specific data service architecture, was addressed in recent years (Amira, Vikas, & Abdelaziz, 2015). There was much research carried out on road accident prevention recently, thus this work is based on road accident data widely available in the public domain.

Association rules mining is one of the best methodologies for extracting features among data in a large dataset and used for frequent itemset mining (Liling, Shrestha, & Hu, 2017). It finds out the relation between the data stored in a large database and provides the significant advantages of determining the frequent patterns. Importantly, this mining technique consumes less memory. Of many association rule algorithms, Apriori algorithm can be implemented easily, but there are many drawbacks. The main drawbacks of the algorithm are obtaining non-interesting rules, a huge number of discovered rules, less algorithm performance, and the least performing for big data in cloud storages/local servers. Thus, association rule alone does not suffice for the RAP miner model to be designed in this work.

Amira et al. (2015) applied association rules mining algorithm on a dataset of traffic accidents which was gathered from Dubai Traffic Office, UAE. Upon information pre-processing, Apriori and Predictive Apriori association rules algorithms were applied for the chosen dataset to investigate the connection between recorded accidents and factors to accident severity in Dubai. Two sets of class association rules were generated using these two algorithms and obtained the most interesting rules using technical measures. The results proved that class association rules created by Apriori algorithm were more viable than those created by Predictive Apriori algorithm. Correlation between accident factors and accident severity level was investigated. While Apriori and Predictive Apriori are useful tools, researchers failed to address the evolving nature of RAPs with respect
to space and time and thus critical pattern fusion techniques are required to extract a matching RAP set. Thus, this will create big data on RAPs over the time. Unless otherwise, researchers address the evolving nature of RAPs interleaving with the repair of existing patterns in the knowledge base already established, it is impossible to predict most frequent accidents in critical places at crucial time intervals.

Zhou and Bao (2008) proposed an algorithm to discover combined association rules to improve the efficiency of prediction. This was done by combining association rule techniques allowing different users to perform actions directly. Rule generation and interestingness measures in combined association rule mining were mainly focused on this work. In combined association rule generation, the frequent itemset among itemset groups was discovered to improve the efficiency. This research failed to address RAP fusion despite the outcomes are useful to eliminate duplicate patterns from a historical pattern set stored in a knowledge base. Hence, the repeatability of the implemented software for new occurrences of RAPs, leading to repair, revise and update the knowledge base, has not been considered.

Gautam and Pardasani (2010) investigated an efficient version of Apriori algorithm for mining multilevel association rules in large databases to find a maximum frequent itemset at the lower level of abstraction. This was done using a fast and efficient algorithm with a single scan of database for mining a complete frequent itemset. The proposed algorithm can derive multiple-level association rules under different supports in a simple and an effective way. Despite single scan pattern mining is fast to retrieve the matching pattern set, accuracy and precision were compromised in this work because no CBR algorithms were introduced in this work. By the inclusion of CBR with a case library with a relevant pattern set isolated from the rest of RAPs, pattern mining, using multilevel association rules, will fail.

Zhou and Bao (2008) proposed an algorithm for double connective association rule mining for which a three-table relational database was used to discover rules for mining. The rules were found among the primary keys of two entity tables and the primary key of binary relationship table. Unless otherwise, a case library is formed using pre-injury (no accident pattern set), injury pattern set (during the occurrence of accidents), and post-injury (rehabilitation of accident patterns using repair and revise of RAPs), double connective association rule mining comprises most recent occurrences of RAPs leading to failure in the prediction.

Basso et al. (2018) investigated the performance of feature selection approach in terms of classification model’s sensitivity and robustness. They proposed a model in which two classification (support vector machine and logistic regression) methods were used for predicting the accident with promising results. These methods yield a high level of sensitivity; however, they provide high false-positive rates, overestimating the zones where accident would probably occur. It seems higher sensitivity does not give accurate prediction rates unless the false-positive rate is minimized. Therefore, it is not feasible to use a model with a high false-positive rate. The overestimating zones were discovered because no case library formed during the evolution of RAP occurrences leading to appropriate pattern fusion.

Janani and Devi (2018) investigated the performance of the feature selection approach in terms of predicting the model’s accuracy. They proposed the system that uses Naive Bayes classification algorithm to predict the severity of the accident which performed
well in terms of accuracy increasing it to 92.45%. Even though with a high accuracy rate from Naïve Bayes classification than other algorithm (Decision tree J48, Random Forest) was reported, it assumes independence among the variables i.e. it does not yield correct estimation when one of the variables has zero occurrence. Therefore, Naïve Bayes classification with a high accuracy rate led to the prediction of accident occurrence if a new variable was presented. While the classification of RAP occurrences is the best outcome in this research, critical features, taking into consideration space and time, have not been addressed using a case library together with CBR. Thus, the need for hybrid intelligent algorithms arises from this work.

Park et al. (2016) investigated the performance of the feature selection approach in terms of predicting the model’s accuracy. They proposed a model, which used oversampling to balance the data set and use the Map Reduced-base method to increase the accuracy to 76.35%. Whilst this is an improvement over the performance shown by SVM and other algorithms which was 73.63%, synthetic data are added to balance the data that makes performance unrealistic. As the data are added to make the imbalance data perform with higher accuracy, other techniques shall be reconsidered to increase the accuracy. RAP mining and pattern fusion were not addressed appropriately in this work leading to synthetic data inclusion for prediction. The need for CBR was not well understood.

Kumar and Toshniwal (2015) performed partition-based clustering and density-based clustering to group similar accidents together to predict the patterns. Based on the categorical nature of most of the data, K-modes algorithm was used to find the correlation among various sets of attributes. The data set was classified into six clusters and each of them was studied to predict some accident patterns. While similar accidents are essential, prior pattern sets and post-pattern sets during pre- and post-accident occurrences with relevant pattern sets are critical requiring pattern fusion.

Krishnaveni and Hemalatha (2011) worked with classification models to predict the severity of injury that occurred during traffic accidents. They used various classifiers such as Naive Bayes Bayesian classifier, AdaBoostM1 Meta classifier, PART Rule classifier, J48 Decision Tree classifier, and Random Forest Tree classifier to compare and classify the type of injury severity of various traffic accidents. The final result showed that the Random Forest outperformed the other four algorithms. A comparative analysis done in this work is useful for case library formation using hybrid intelligent algorithms that lead to optimal and semi-optimal solutions.

Kashani and Mohaymany (2011) investigated classification accuracy on the basis of performance metric to predict the severity of accidents. They proposed a model, which used the Classification and Regression Tree (CART) to analyze road accidents data in Iran and found that not using seat belt, improper overtaking and over speed affected the severity of accidents. Unless otherwise, CART addresses real-time RAP occurrences together with case library, the decision fusion reached in this work will have no significance in terms of the prediction accuracy and repeatability.

Preeti, Gupta, Singh, and Dhiman (2017) investigated the performance of prediction approach in terms of predicting the model’s accuracy by applying association rule mining algorithms.

Chen and Jovanis (2000) investigated the performance of feature selection for predicting the model’s accuracy. This was done by analyzing large dimensional datasets using
traditional statistical techniques in large contingency tables and the statistical models with specific assumptions and violation using data mining techniques to analyze road accidents. Data mining was used to extract novel, implicit and hidden information from large data. Feature selection and data mining are useful tools for RAP occurrences as far as appropriate case library is formed and evolved.

Castro et al. (2016) investigated the classification accuracy of multilevel classification using classified documents. They used WEKA’s Bayes net classifier and achieved an accuracy of 72.39% with precision on 0.725, 0.212 and 0.466 for slight, serious, and fatal, respectively. Bayes Net Classifier was the most accurate model against J48 decision tree and Multi-Layer Perceptron (MLP) model, which had the accuracy of 72.02% and 71.70%, respectively; but it did not consider all factors and interaction among them for prediction led to unstable classifier. It is important to consider the pattern fusion taking into consideration CBR by revising, reusing, and retaining the existing pattern in a RAP case library. This will improve, update, and modifying the classification subject to pre-injury (no accident), injury (accidents), and post-injury (rehabilitation) pattern set from the road accident patter miner already established than just updating a knowledge base using new patterns.

Hasheminejad (2017) investigated classification accuracy based on a performance metric. A novel multi-object and rule-based method outperformed the classification methods, such as ANN, SVM, and conventional DTs according to classification metrics’ accuracy (88.2%) and performance metrics of rules; support and confidence (0.79 and 0.74, respectively). The proposed method yielded promising result with an increased accuracy of 4.5% with respect to other methods but the obtained rules from this method were not effective. Therefore, the feature selection method and extraction method shall be used to increase the accuracy and improve the effectiveness of the obtained rules. Establishing a road accident case library through a transformed RAP database allows removal of some patterns set not influencing the classification helps in extracting a matching pattern set from the case library already established using CBR. Thus, the optimal reuse of the existing pattern set from a case library yields the accuracy and effectiveness of classification.

Gu et al. (2017) investigated particle swarm optimization, which demonstrated prediction with more accuracy and precision. Particle swarm partition had the accuracy with high precision compared to mutation optimization back propagation neural network prediction model, particle swarm optimization-support vector machine model, support vector machine, back propagation neural network, K Nearest Neighbour (K-NN), and Bayesian network. But it generated an unknown factor which was avoided. The drawback of particle swarm optimization is to trap into a local optimum which shall be avoided using an appropriate transformed road accident database resulting in revising, reusing, retaining new cases for the case library using CBR. This will allow extracting the best matching pattern set using the optimal parameter set already classified in the improved, updated, and repaired case library.

**RAP model using Naïve Bayes classification and Apriori algorithm**

Janani and Devi (2018) proposed Naïve Bayes classification technique using Apriori algorithm association mining to predict the severity of accident. Even though Janani and Devi
(2018) introduced a classification model, it failed to consider revising, reusing, and retaining RAP set taking into consideration CBR for evolving the nature of knowledge base. Thus, in this work, the concept of RAP Database (RAP DB), knowledge base and the creation of Case Library (CL) has been introduced as a hybrid learning method. Therefore, retrieving the matching pattern set from the RAP DB formed is done for actual RAP set under consideration as a repetitive process for evolving CL. The prediction of location is based on fuzzy inference engine (FIE) built by retrieving a relevant accident frequency pattern from RAP DB, as illustrated in Figure 1. Naïve Bayes Classifier has improved the accuracy of prediction model by Janani and Devi (2018). RAP DB was implemented by clustering a RAP set generating the rules based on Apriori algorithm. RAP stored in the database is then decomposed into a set of training RAP set and test RAP set (Splitting RAP DB) to predict the severity of accident. The results demonstrated an accuracy of 92.45% in predicting accident. Thus, the RAP model leading to RAP Case Library (RAP CL) was developed, as illustrated in Figure 1 using five processes, i.e. RAP Pre-processing, Clustering, Association rules mining, Classification, and Prediction using CBR. This will allow repairing/updating actual RAP, reusing existing RAP from RAP DB, and retaining as a new RAP set presented for RAP CL if the actual RAP set is not available in the existing RAP CL.

**RAP pre-processing**

RAP pre-processing starts with balancing the RAP set in the RAP database (RAP DB) as illustrated in Figure 1. RAP set in RAP DB goes through three fundamental operations: duplication, missing attributes of a RAP (more than 50% zeros in a RAP) and incomplete pattern due to less than 50% zeros in a RAP which were similarly carried out in Joshi et al. (2020), but in this work, it is done differently. According to the first operation for a RAP set,
as illustrated in Figure 1, if a duplication RAP exists, it is removed from the RAP DB, otherwise subsequently checked for a RAP with more than 50% zeros in its attributes. If a RAP is found, it is also removed from the RAP DB. As the last process of RAP pre-processing, if a RAP with less than 50% zeros in its attributes is found, the corresponding pattern is replaced with the attribute mean value calculated from the raw RAP set in RAP DB. These three processes are repeated through all RAP set within RAP DB. Thus, the transformed RAP DB is formed from the RAP pre-processing, as illustrated in Figure 1.

**RAP clustering**

During the RAP clustering process, the RAP set in the transformed RAP DB (Figure 1) is grouped in such a way that the attributes of each RAP in one cluster differ from another cluster. Based on the accident locations extracted, three different clusters, $C_1$, $C_2$, and $C_3$ as shown in Figure 1, are formed using the locations grouping so that each RAP within the same RAP cluster shows the same behaviour or each RAP cluster contains RAPs with the same attributes. This clustering was done using K-means algorithm. Once the clustering is done, the association rules mining technique using Apriori algorithm is implemented to identify the set of rules as proposed by Janani and Devi (2018) and as shown in Figure 1. The extraction of rules or relationships from RAPs discovers that many times the same or similar RAP set has been recorded over a period. Identification of rules between different clusters and RAP set can be used to classify the level. The solution proposed by Janani and Devi (2018) generated frequent RAP set and strong association rules which satisfies the minimum threshold lift with the 5% correlation revealing among different attributes from each RAP of transformed RAP DB.

High prediction accuracy results in higher time consumption, as reported in Janani and Devi (2018), due to the use of Apriori algorithm for feature extraction from the RAP CL formed. The main drawback of RAP CL is the excessive number of RAP set stored in the RAP CL, leading to a large number of RAP sets found on frequent RAP set or which has a low support count. For example, if there are a large number of RAP sets from frequent-1 RAP set, then it needs to generate more RAP sets than a large number of RAP sets into $2^{\text{length}}$ which may be tested and accumulated. To detect the pattern from RAP CL using frequent RAP set of large size $100 \,(P_{1}..P_{100})$, it is required to generate $2^{100}$ RAP sets which may be time consuming and costly for the Apriori algorithm’s RAP generation. Thus, finding candidate RAP sets requires scanning RAP CL many times repeatedly.

**Naïve Bayes classification for the clustered RAP**

The Naïve Bayes Algorithm is used to classify the severity level of accident. Having clustered the transformed RAP DB, rules were generated and decomposed into training RAP set and test RAP set 70% and 30%, respectively to create the splitting RAP DB, as suggested by Janani and Devi (2018). Based on the attributes defined Fatal, Grievous, Damage, and Injury in the transformed RAP DB, class labels are created in which the severity level of accident is reflected. Two class labels are defined in this work; Class 0 represents low severity level and class 1 represents high severity level. Naïve Bayes classifier trained learnt classifiers ($LC_i$) in the splitting RAP DB with the best accuracy, as
illustrated in Figure 1. Based on the classification done, RAP CL enriches RAP set with the factors behind road accidents occurred.

**Fuzzy inference Engine for the classified RAP**

Having done the clustering based on the location of the accident and classification using the severity of the accident, the probability of accident occurrence is determined by applying FIE subject to identifying new cases of accident occurrence by enriching the knowledge base, creating RAP CL (RAP CL), as illustrated in Figure 1. The fuzzification transforms crisp input values into fuzzy values. These fuzzy values are processed using the rule base established in the FIE. The output is obtained by transforming processed fuzzy values into crisp domain using defuzzification, as reported in Joshi et al. (2020). But, in this work case repository is considered such a way that the evolving cases of accident occurrences are populated progressively to establish a knowledge base which is called RAP CL, as illustrated in Figure 1. Thus, the probability of accident occurrences as the output is governed by RAP CL interfaced with FIE compared to two different processes. The generation of accident occurrence cases is based on two different RAPs: initially when the RAP CL is empty and subsequently when the case repository is already populated with some cases of RAP. In the first situation when there is no RAP case in the repository, RAP sets collected and clustered are to form the groups based on their classification done. Hence, the labels of these groups are identified during the classification process. Then new cases are generated based on a semi-automatic process where both extracted features, using transformed RAP DB and recommendation from classified RAP DB, are stored using the proposed case representation. In the later process, a new case is added based on the outcome of CBR cycle that consists in retrieving, reusing, revising and retaining cases in the RAP CL. Hence, re-grouping/re-learning of the classifier is done if required depending on the availability of new case presented to the FIE during the retrieval stage of CBR from RAP CL. In each solved case, RAP was stored (revised and retaind) in the RAP CL, including the description of accident occurred, mean value of transformed attribute/value pairs for clustering and classification during a new RAP case of accident presented to the FIE and RAP CL. Thus, FIE and RAP CL execute interactively to enrich the RAP CL in such a way that prediction accuracy increases during progressive updates on the RAP CL subject to a new occurrence of accident identified by the FIE based on the RAP miner already built and progressing. Thus, RAP Miner allows the prediction of accident occurrence using the probability of accident occurrences, which provides feedback for road accident prevention authorities in the community.

**Mathematical modelling for the RAP miner**

The Apriori Algorithm is implemented for association rules mining to extract a set of rules, which defines a RAP set, as suggested by Janani and Devi (2018). However, the processing time can be further be reduced to extract rules based on the considerable decrease of overall processing time.

Let \( l = \{i_1, i_2 \ldots i_n \} \) denote the set of n RAP, and let \( D = \{t_1, t_2 \ldots t_m \} \) denote the set of transactions;

Association rule shall be defined as follows.
An implication of form of $A \rightarrow B$ such that $A, B \subset I$ and $A \cap B = \phi$. Extracting rules that provide the correlation between the attributes when accident happens, based on the Lift value given by Janani and Devi (2018). Support for the association rule $A \rightarrow B$ is the percentage of transaction in the RAP DB that contains $A \cup B$, as shown in Equation (1) proposed by Janani and Devi (2018):

$$Supp(A \rightarrow B) = Supp(A \cup B) = \frac{|t(A \cup B)|}{t(A)}$$

Confidence for the association rule is $A \rightarrow B$ is the ratio of the number of transactions that contains $A \cup B$ to the number of transactions that contain $A$, as shown in Equation (2) proposed by Janani and Devi (2018):

$$Confidence(A \rightarrow B) = \frac{Supp(A \cup B)}{Supp(A)}$$

To calculate the processing time complexity of Apriori Algorithm for feature extraction to decrease the processing time for the prediction accuracy of accident occurrence, Janani and Devi (2018) proposed Equation (3):

$$T = \sum_k (2E(|x|)) + m_k k \tau$$

where

- $n$ is the number of features from a RAP.
- $k$ is the length.
- $\tau$ is the time per non-zero element.
- $E(|x|)$ is the number of times features extracted from RAP for the mean complexity value.
- $m_k k$ is to the complexity of the search space which is visited for each RAP.

### Scan Efficient Apriori (SEA) algorithm for the RAP miner built

SEA generates accident cases for RAP CL as a new RAP case to retain faster causing considerable time reduction during the processing of RAP DB. This algorithm reduces the number of RAP set to be scanned in RAP DB due to a significant number of similar cases identified from the RAP CL already formed. Whenever the $k$ of $k$-RAP set and the

| Table 1. Apriori Algorithm for RAP miner. |
|-----------------------------------------|
| **Algorithm:** | Apriori algorithm to extract rules to identify RAP |
| **Begin** | Ck: Candidate RAP set of size k |
| | Lk: frequent RAP set of size k |
| | L1 = {frequent RAP}; |
| | for (k = 1; L k= \phi k++) |
| | do begin |
| | Ck + 1 = candidates generated from L k; |
| | for each transaction t in RAP CL increment the count of all candidates in Ck + 1 that are contained in t |
| | Lk + 1 = candidates in Ck + 1 with min. support |
| | end |
| | return Uk Lk; |
value of minimum support increase based on the Apriori Algorithm shown in Table 1, the SEA algorithm scans RAP miner faster, as reported by Arwa and Mourad (2018). Thus, the breadth-first search algorithm exploits the downward closure property of support to count the RAP set together with appropriate candidate case generation for RAP CL which is one of the best ways to mine the RAP/association rules. SEA acts in a way minimizing the scan in RAP DB and as well as in the transformed RAP DB making it possible to optimize the generation of candidate RAP for RAP CL going through CBR stages: retrieve, reuse, revise, and retain. The SEA is implemented using new transaction mapping which avoids the repetitive scan of RAP miner, improving the joining efficiency by pruning two RAP sets (frequent and candidate) and achieving higher efficiency by reusing existing RAP from RAP CL. Hence, the results illustrate that a novel SEA algorithm is the optimal solution with a significant decrease of processing time compared to the rest of algorithms based on the similar comparisons reported in Joshi et al. (2020).

Thus, the overall processing time required shall be calculated by introducing time $t_f$ required for scanning RAP DB for the first time as follows:

$$T = t_f \times m_k + \sum_{i=1}^{n} (2E(|x|)) + k) \tau$$

(4)

where

- $t_f$ is the time required for the first time scanning of RAP DB.
- $m_k$ is the complexity of search space for RAP set $k$.
- $k$ is the length of RAP set.
- $E$ is the feature extracted from RAP for the mean complexity value.
- $|x|$ is the mean complexity of each RAP.
- $\tau$ is the amount of time per non-zero element.

By implementing the SEA algorithm, it scans RAP DB to generate frequently occurring first RAP set which is then used to find the second RAP set and so on until $k$ number of RAP set is reached. Thus, it minimizes processing time as previously stored RAP case is reused from RAP CL if RAP exists, otherwise revises a new RAP case under consideration causing it to revise and to retain the new case leading to re-learning RAP CL, as illustrated in Table 2.

Table 2. SEA Algorithm for retaining new RAP case in RAP CL.

| Algorithm: SEA algorithm to extract rules to identify RAP |
| Begin |
| Step (1) first RAP set $L_k$ = find frequent 1 RAP set ($T$); |
| Step (2) For loop $k$ from 2 to while $L_k-1$! = null; |
| //Generate the Candidate RAP set from the $L_k-1$ |
| Step (3) $C_k$ = candidates generated from $L_1$ |
| //get the RAP with minimum support in candidate RAP set using $L_1$ ($1 \leq w \leq k$). |
| Step (4) $x$ = Get RAP set minimum support($C_k$, $L_1$); |
| // get the target transaction ID (TID)s that contain RAP $x$. |
| Step (5) $TID$ = get Transaction ID($x$); |
| Step (6) For each transaction $t$ in $TID$ Do |
| Step (7) Increment the count of RAP in candidate RAP found in $TID$; |
| Step (8) $L_k$ = RAP set in candidate RAP set $\geq$ minimum support; |
| End; |
Therefore, the introduction of SEA algorithm shall be reflected in the RAP miner, as illustrated in Figure 2 (inside the green box).

**Analysis of results for RAP miner built**

Anaconda navigator, Spyder and scikit-learn are used to implement SEA algorithm for RAP Miner. RAP DB is created using open access site data.gov. RAP set available on the website is in CSV files that are used to create RAP DB for clustering purpose. The total 4500 RAP set with a maximum of 300 features from which significantly influential features, used for the RAP miner such as date and time of the accident case, location, and severity of the accident, etc., are included.

SEA algorithm is implemented for the chosen RAP set in RAP DB formed using Equation (4). Table 3 illustrates the processing time and accuracy of the features extracted from RAP miner for different RAP cases stored during the formation of RAP DB. RAP case generation for RAP CL to be created begins with the first time scanning the transformed DB. Thus, the

| RAP DB from data.gov                  | RAP sample set (n) | Candidate RAP set (m_q) | Apriori algorithm Process. time | Accuracy | SEA Process. time | Accuracy |
|---------------------------------------|--------------------|-------------------------|--------------------------------|----------|------------------|----------|
| RAP set                               | 915                | 62                      | 6.001                          | 87.23    | 2.119            | 90.33    |
| New crash Statistics                  | 1625               | 135                     | 6.986                          | 89.25    | 2.706            | 90.27    |
| Crash and RAP                        | 4908               | 275                     | 62.767                         | 81.67    | 15.175           | 85.61    |
| Pedestrian crashes                    | 3578               | 248                     | 31.665                         | 85.66    | 8.762            | 91.34    |
| Research group data for RAP          | 572                | 45                      | 1.717                          | 82.45    | 0.613            | 86.77    |
| Annual crash fact sheet              | 2378               | 186                     | 10.001                         | 87.02    | 2.997            | 91.07    |
| Road crash RAP                       | 1122               | 98                      | 9.455                          | 92.45    | 4.134            | 92.75    |
| Fatal, serious and all serious crash | 1866               | 157                     | 7.676                          | 80.68    | 2.865            | 89.87    |
| RAP by time of occurrence            | 787                | 60                      | 3.655                          | 85.79    | 1.342            | 91.22    |
| Crash by location                    | 2680               | 201                     | 13.418                         | 89.78    | 3.039            | 92.02    |
average processing time is reported based on different RAP sets obtained from data.gov, as illustrated in Table 3.

RAP miner in Table 3 demonstrates that processing time (Process. Time) and accuracy perform far better in SEA algorithm compared to the traditional Apriori algorithm. Thus, the formation of RAP CL in RAP miner started from the first scan of RAP DB, starts searching for a reduced number of RAP sets in the subsequent scan, facilitates fast processing and accurate feature extraction for classification and for FIE compared to scanning whole RAP DB by Apriori algorithm in each scan of RAP DB with no RAP CL created during the first scan of RAP DB.

Based on the minimum value support for the association rule mining, as discussed in Equations (1) and (2) during the transaction through transformed RAP DB, RAP miner results are analyzed for the same RAP DB formed, as illustrated in Table 4. Different RAP sets from different sources from data.gov are taken into consideration for this analysis. Varying number of RAP sample sets and its candidate RAP sets are considered for classification and FIE to enrich RAP CL. The results shown in Table 4 prove that the association rules using SEA for different minimum values support are outperformed. Hence, SEA provides the optimal solution for RAP miner by creating RAP CL progressively based on new RAP case to be generated. SEA does not scan through the complete RAP DB, rather it looks for reusing RAP cases from the established RAP miner using its RAP CL. If a matching RAP case is available no new RAP is generated, otherwise, a new RAP case is generated for retaining a new RAP case in the RAP CL using re-learning of RAP CL. Thus, RAP CL is formed with the least search through RAP DB resulted in fast RAP set processing.

### Table 4. Processing time and accuracy for feature extraction with minimum value of support.

| RAP DB from data.gov | RAP sample set (n) | Candidate RAP set (m) | Min support | Apriori Algorithm | SEA with RAP miner |
|---------------------|--------------------|-----------------------|-------------|-------------------|-------------------|
|                     | Process. time | Accuracy | Process. time | Accuracy |
| Research group RAP  | 572              | 45        | 0.02         | 4.776       | 82.67     | 0.877       | 89.21 |
| RAP set             | 1122             | 98        | 0.02         | 4.852       | 86.77     | 0.942       | 89.56 |
| New crash statistics| 1625             | 135       | 0.02         | 5.011       | 81.54     | 0.995       | 87.69 |
| Crash by location   | 2680             | 201       | 0.02         | 5.669       | 82.09     | 1.011       | 86.67 |
| Pedestrian crashes  | 3578             | 248       | 0.02         | 6.638       | 77.67     | 1.056       | 85.05 |
| Research group RAP  | 572              | 45        | 0.04         | 0.762       | 80.60     | 0.314       | 89.33 |
| RAP set             | 1122             | 98        | 0.04         | 0.811       | 86.79     | 0.356       | 91.02 |
| New crash statistics| 1625             | 135       | 0.04         | 0.927       | 84.62     | 0.398       | 87.24 |
| Crash by location   | 2680             | 201       | 0.04         | 1.034       | 79.78     | 0.412       | 81.62 |
| Pedestrian crashes  | 3578             | 248       | 0.04         | 1.855       | 75.01     | 0.422       | 79.83 |
| Research group RAP  | 572              | 45        | 0.06         | 0.576       | 72.17     | 0.177       | 81.88 |
| RAP set             | 1122             | 98        | 0.06         | 0.734       | 80.64     | 0.213       | 91.53 |
| New crash statistics| 1625             | 135       | 0.06         | 0.956       | 75.05     | 0.259       | 85.47 |
| Crash by location   | 2680             | 201       | 0.06         | 1.042       | 90.35     | 0.298       | 90.67 |
| Pedestrian crashes  | 3578             | 248       | 0.06         | 1.158       | 78.52     | 0.330       | 86.45 |
| Research group RAP  | 572              | 45        | 0.08         | 0.212       | 72.87     | 0.128       | 81.77 |
| RAP set             | 1122             | 98        | 0.08         | 0.278       | 85.93     | 0.149       | 87.82 |
| New crash Statistics| 1625             | 135       | 0.08         | 0.336       | 79.54     | 0.165       | 81.05 |
| Crash by location   | 2680             | 201       | 0.08         | 0.389       | 81.06     | 0.178       | 82.34 |
| Pedestrian crashes  | 3578             | 248       | 0.08         | 0.424       | 80.65     | 0.199       | 89.01 |
| Research group RAP  | 572              | 45        | 0.10         | 0.198       | 72.97     | 0.105       | 83.32 |
| RAP set             | 1122             | 98        | 0.10         | 0.237       | 77.77     | 0.119       | 81.67 |
| New crash statistics| 1625             | 135       | 0.10         | 0.275       | 84.33     | 0.138       | 90.65 |
| Crash by location   | 2680             | 201       | 0.10         | 0.319       | 82.09     | 0.152       | 83.76 |
| Pedestrian crashes  | 3578             | 248       | 0.10         | 0.382       | 71.67     | 0.168       | 89.34 |
**Table 5.** Comparative results of existing solution and novel solution.

|                        | Novel solution                                                                 | Existing solution                                                                 |
|------------------------|--------------------------------------------------------------------------------|----------------------------------------------------------------------------------|
| Function               | SEA Algorithm interfaced with RAP CL                                          | Association rules mining using Apriori Algorithm                                  |
| Accuracy               | 92.75% with optimal processing time                                           | 92.45% during prediction with increased processing time                           |
| Processing time        | Decrease with minimum support 0.2 = 1.056 s 0.4 = 0.422 s 0.6 = 0.330 s 0.8 = 0.199 s 0.10 = 0.168 s Decreased by 67.38% | For the same minimum support 0.2 = 6.638 s 0.4 = 1.855 s 0.6 = 1.158 s 0.8 = 0.424 s 0.10 = 0.382 s |
| Proposed equation      | Processing time complexity                                                      | Processing time complexity                                                        |
|                        | $T = t_1 \times m_k + \sum_{i=1}^n (2E(|x|)) + k)\tau$                      | $T = \sum_k (2E(|x|)) + m_kn\tau$                                                |
| Contribution           | Performs efficiently as it finds the frequent occurring RAP for the RAP CL to build from the first scan for new case generation to retain in the RAP CL such that during the second scan retain is based on CBR | Performs efficiently when the number of transaction increases leading to repetitive scan through the complete RAP DB. Computational resources grow with the increase in RAP DB cases |
| Key advantage          | Avoids repetitive scan of complete RAP DB by creating RAP CL using hybrid learning | Scans the RAP DB continuously making overall performance poor                      |

**Discussion**

RAP miner is built by the formation of RAP CL using a novel hybrid learning algorithm interfaced with SEA algorithm for association rules mining such a way that first time RAP DB scan assures the formation of RAP CL paving way for subsequent RAP case generation and retention subject to CBR. Thus, SEA with CBR improved the overall performance of RAP miner with an accuracy of 92.75% compared to the maximum accuracy of 92.25% obtained from the rest of the algorithms published. Due to the least number of transactions through RAP DB, the processing time is reduced by 67% which is a remarkable achievement of the performance in the RAP miner. Table 5 shows the key advantages of SEA algorithm interfaced with RAP CL for the RAP miner built in this work.

**Conclusion and future directions**

The RAP miner is successfully built and validated using an open-source RAP DB widely available online. A novel hybrid learning algorithm interfaced with SEA algorithm outperforms compared to the results reported due to the formation of RAP CL for new RAP case presented using CBR that has not been reported until today.

As future directions, the RAP miner built can predict RAP based on the location and severity of the accident as feedback for drivers in real-time, perhaps by avoiding clustering, based solely on the SEA algorithm built using reinforcement learning interfaced with CBR.

**Disclosure statement**

No potential conflict of interest was reported by the author(s).

**Notes on contributors**

S. M. N. Arosha Senanayake is in academia within four THE and QS ranking universities in the world achieving excellence in research, academic, and administrative portfolios. He has already published more than 175 scholarly articles in high impact factor journals, commercial book chapter
publications, books, and proceedings of peer reviewed conferences. Most recently his team was the recipient of research excellence award as the best project award among 6 leading projects within ASEAN+ (ASEAN + Japan) during 2017-2020. Arosha was invited to set up IntelliHealth Solutions (Technology Licensing) in 2015 and he continues to thrive projects as the founder and leader of IntelliHealth Solutions jointly with National Institute of Information and Communications Technology, Japan, Gifu University, Japan, University of Malaya Connected Health Pvt Ltd, Malaysia, and other multi-national industries around the world.

Sisir Joshi was a Master’s graduate from Charles Sturt University of Technology during which he introduced novel algorithms for this project and published his work as a book chapter. In this article, his algorithms implemented during master’s were enhanced using hybrid intelligent algorithms and pattern miner which is completely novel in road accidents research.

References

Amira, A., Vikas, P., & Abdelaziz, A. (2015). Applying association rules mining algorithms for traffic accidents in dubai. International Journal of Soft Computing and Engineering, 5(4), 2231–2307. http://www.ijscce.org/wpcontent/uploads/papers/v5i4/D2679095415.pdf

Arwa, A., & Mourad, Y. (2018). I hybrid approach for improving efficiency of apriori algorithm on frequent itemset. International Journal of Computer Science and Network Security, 18(5), 7–10. http://paper.jicnsn.org_07_book_201805/20180521.pdf

Basso, F., Basso, L. J., Bravo, F., & Pezoa, R. (2018). Real-time crash prediction in an urban expressway using disaggregated data. Transportation Research Part C: Emerging Technologies, 86, 202–219. https://doi.org/10.1016/j.trc.2017.11.014

Castro, F. M., Marin-Jiménez, M. J., & Guil, N. (2016). Multimodal features fusion for gait, gender and shoes recognition. Machine Vision and Applications, 27(8), 1213–1228. https://doi.org/10.1007/s00138-016-0767-5

Chen, W., & Jovanis, P. (2000). Method for identifying factors contributing to driver-injury severity in traffic crashes. Transportation Research Record: Journal Of The Transportation Research Board, 1717 (1), 1–9. https://doi.org/10.3141/1717-01

Gautam, P., & Pardasani, K. (2010). A fast algorithm for mining multilevel association rule based on boolean matrix (IJCSE). International Journal on Computer Science and Engineering, 2(3), 746–752.

Gu, X., Li, T., Wang, Y., Zhang, L., Wang, Y., & Yao, J. (2017). Traffic fatalities prediction using support vector machine with hybrid particle swarm optimization. Journal Of Algorithms & Computational Technology, 12(1), 20–29. https://doi.org/10.1177/1748301817729953.

Hasheminejad S. M. H. (2017). Traffic accident severity prediction using a novel multi-objective genetic algorithm. International Journal Of Crashworthiness, 22(4), 425–440. https://doi.org/10.1080/13588265.2016.1275431

Jana, G., & Devi, N. (2018). Road traffic accidents analysis using data mining techniques. JITA - Journal Of Information Technology And Applications (Banja Luka) - APEIRON, 14(2), 84–91. https://doi.org/10.7251/JIT1702084J

Joshi, S., Alsadoon, A., Arosha Senanayake, S. M. N., Prasad, P. W. C., Yin Yong, S., Elchouemi, A., & Hung Vo, T. (2020). Pattern mining predictor system for road accidents. In M. Hernes, K. Wojtkiewicz, & E. Szcerbicki (Eds.), Advances in Computational collective intelligence. ICCCI 2020. Communications in Computer and information science (Vol. 1287, pp. 605–615). Springer. https://doi.org/10.1007/978-3-030-63119-2_49.

Kashani, A. T., & Mohaymany, A. S. (2011). Analysis of the traffic injury severity on two-lane, two-way rural roads based on classification tree models. Safety Science, 49(10), 1314–1320. https://doi.org/10.1016/j.ssci.2011.04.019

Krishnaveni, S., & Hemalatha, M. (2011). A perspective analysis of traffic accident using data mining techniques. International Journal of Computer Applications, 23(7), 40–48. https://doi.org/10.5120/2896-3788
Kumar, S., & Toshniwal, D. (2015, December 4–5). Analysing road accident data using association rule mining. 2015 International Conference on Computing, Communication and Security (ICCCS), Pointe aux Piments, Mauritius, pp. 1–6. https://doi.org/10.1109/cccs.2015.7374211

Liling, L., Shrestha, S., & Hu, G. (2017, June 7–9). Analysis of road traffic fatal accidents using data mining techniques. 2017 IEEE 15th International Conference on Software Engineering Research, Management and Applications (SERA), London, UK, pp. 363–370. https://doi.org/10.1109/SERA.2017.7965753

Park, S., Kim, S., & Ha, Y. (2016). Erratum to: Highway traffic accident prediction using VDS big data analysis. The Journal Of Supercomputing, 72(7), 2832–2832. https://doi.org/10.1007/s11227-016-1655-5

Preeti, B., Gupta, S., Singh, P., & Dhiman P. (2017, May 5–6). Accident and road quality assessment using android google maps API. 2017 International Conference on Computing, Communication and automation (ICCCA), Greater Noida, India, pp. 1061–1064. https://doi.org/10.1109/CCAA.2017.8229952

Zhou, X., & Bao, H. (2008, December 12–14). Mining double-connective association rules from multiple tables of relational databases. 2008 International Conference on Computer Science and Software Engineering, IEEE Computer Society, Wuhan, China. https://doi.org/10.1109/csse.2008.559