ISSN 2349-4506
Impact Factor: 3.799

Global Journal of Engineering Science and Research Management

RARE SEVERE PATTERNS IN ROAD ACCIDENT RECORDS: A REVIEW

Desai Saurabh Sudhir[1], Arpit Solanki[2]
[1]Dr. A.P.J. Abdul Kalam University, Indore, M.P., India
[2]Dr. A.P.J. Abdul Kalam University, Indore, M.P., India

KEYWORDS: Association rule mining; frequent pattern mining; FP-Growth; class association rules; road traffic accidents; crash severity; interestingness measures; false discovery rate.

ABSTRACT
Road traffic crashes remain a major public-health burden, and the police records that describe them are a rich but under-used source of safety knowledge. Association rule mining turns such records into readable if–then statements, which makes it attractive for road-safety work. This paper reviews the literature that bears on applying it to crash data: the foundations of frequent pattern mining and the efficiency techniques built on them, condensed representations, multiple-support and weighted mining, class association rules, interestingness measures, class imbalance, the statistical significance of discovered patterns, and data mining studies of road accidents. The surveyed methods are then compared in terms of what each contributes and what limits it for severity-labelled crash records, with attention to weighting, class-specific support thresholds, class-targeted mining and removal of redundant rules. The comparison shows that each property has been studied in isolation, mostly on generic benchmark data, whereas road-safety applications generally use existing miners without changing them. No reviewed method combines severity-aware weighting, class-specific thresholds, class-targeted search, redundancy control and protection against chance rules, and this absence is identified as the research gap. Based on this gap, the paper points to one possible direction for future work.

INTRODUCTION
Road transport underpins economic activity and everyday mobility, yet crashes on the road kill more than a million people every year and injure tens of millions more, and the burden falls most heavily on low- and middle-income countries [1]. Police-recorded crash statistics compiled annually by the Indian transport ministry likewise show hundreds of thousands of crashes and a very large number of deaths [2]. Each underlying record notes the time and place of the crash, the weather, light and road conditions, the vehicles involved and how severe the outcome was, so these databases hold far more information than summary tables and counts can reveal.

Extracting knowledge from such records is an instance of data mining, the search for previously unknown, valid and useful patterns in large collections of data, which is itself one stage of the wider knowledge discovery process [3]. Among the available techniques, association rules are particularly convenient for safety practitioners because they are stated as simple if–then implications that can be read, questioned and acted upon without statistical training [4]. Accident records, however, are not retail baskets: every record carries a severity label, severe outcomes are rare, and a fatal crash matters far more than a slight one. A miner that uses one global support threshold and counts all records equally therefore tends to lose the patterns of rare severe outcomes [5], [6] and to return more rules than a practitioner can read.

The purpose of this paper is to review the literature around this task and to identify, from the evidence, the research gap that remains open. The review is thematic and selective and rests on 24 sources covering the foundations of frequent pattern mining, condensed, weighted and class-oriented mining, interestingness and significance, and data mining for road safety. Section II surveys this literature, Section III compares the surveyed work and formulates the research gap, and Section IV concludes and, within the future scope, points to one possible solution.

LITERATURE SURVEY
2.1 Foundations of frequent pattern mining
The idea of mining association rules was introduced for customer purchase data, where support and confidence were proposed as the two measures of a rule and the task was defined as finding every rule that meets user-given thresholds [4]. The first efficient solution, Apriori, explores itemsets level by level and discards any candidate that has an infrequent subset, relying on the fact that support can only fall as an itemset grows [7]. Apriori became the reference point for later work, and an early survey and comparison of the competing algorithms was prepared by

Hipp et al. [8]. Repeated database scans and very large candidate sets on dense data led Han et al. to the pattern-growth approach, in which the database is compressed into a prefix tree called the FP-tree and frequent itemsets are obtained recursively without generating candidates [9]. Fournier-Viger et al. later compiled a broad survey of itemset mining that covers frequent, closed, maximal, rare, weighted and high-utility patterns [10].

2.2 Efficiency and condensed representations

Later work reduced the cost of mining and the volume of its output in different ways. Vertical layouts store, for each item, the list of transactions that contain it and obtain the support of an itemset by intersecting such lists; Zaki and Gouda added diffsets, which keep only the difference between the lists of an itemset and its prefix and so save memory on dense data [11]. On dense data the number of frequent itemsets can also grow so large that the output itself becomes the bottleneck. A closed itemset is a frequent itemset with no proper superset of equal support; Pasquier et al. showed that closed itemsets form a lossless summary from which all frequent itemsets and their rules can be derived [12], and Wang et al. mined them efficiently inside an FP-tree framework with CLOSET+ [13].

2.3 Multiple supports, weights and class association rules

A second group of studies makes the mining task fit what the user needs. A single minimum support treats all items alike and forces an awkward compromise, since a high value loses rare but important items and a low value leads to a combinatorial explosion. Liu et al. therefore assigned a separate minimum support to every item and kept a form of downward closure by ordering the items according to their thresholds [5]. Weighted association rule mining goes further by giving importance values to items or transactions; Tao et al. defined a weighted support together with a significance framework that preserves a usable pruning property [14].

When the data carry a class label, the interesting rules are those whose consequent is a class value. Liu et al. combined classification with rule mining in the CBA system, which mines such rules, orders them by confidence and support, and labels a new instance with the first rule that covers it [15]. Yin and Han later proposed CPAR, which builds predictive rules greedily [16]. Together these works indicate that rule-based classifiers can reach the accuracy of conventional classifiers while staying easy to interpret.

2.4 Interestingness, class imbalance and significance

Support and confidence alone let many uninteresting rules through, because a rule can have high confidence simply by virtue of a common consequent. Geng and Hamilton surveyed interestingness measures together with criteria such as conciseness, coverage, reliability and novelty [17].

Two further issues matter when the class of interest is rare. When one class is much smaller than the others, learning methods tend to favour the larger one, and He and Garcia reviewed the remedies, which include resampling, cost-sensitive learning and adjustments at the algorithm level [6]. The second issue is chance. Mining at low thresholds tests a huge number of candidate rules at once, so some of them appear to be associated with the outcome by luck. Webb treated pattern discovery as a multiple hypothesis testing problem and showed that, without correction, a large share of the reported patterns can be false discoveries, and he proposed holdout and direct-adjustment procedures that limit the risk of reporting any spurious pattern [18]. Benjamini and Hochberg introduced the false discovery rate, the expected share of false findings among all findings, together with a step-up procedure that has more power than family-wise corrections such as Bonferroni [19].

2.5 Data mining for road traffic safety

Association rules have been applied to road-safety data in several settings. Geurts et al. used them to profile high-frequency accident locations in Belgium and found combinations of circumstances that set hazardous locations apart [20]. Pande and Abdel-Aty applied market-basket analysis to crash data from large jurisdictions and proposed it as a decision-support tool for safety engineers [21]. Montella related contributory factors to crash types at urban roundabouts through association rules [22], and Kumar and Toshniwal combined clustering with association rules to characterise accident-prone locations on Indian roads [23]. Statistical models of crash injury severity remain the main analytical tools in the field, and the methodological alternatives were reviewed by Savolainen et al. [24].

ISSN 2349-4506
Impact Factor: 3.799

Global Journal of Engineering Science and Research Management

COMPARATIVE ANALYSIS

The surveyed studies address different parts of the mining pipeline and were designed for different data and purposes, so their results should not be compared number by number. The comparison below therefore focuses on what each contribution offers and on the limitation that matters for severity-labelled accident records.

3.1 Comparison of the reviewed studies

Table 3.1 summarises the main studies and the limitation of each that is relevant to this review.

Study	Technique	Main contribution	Important consideration
Agrawal et al. [4]; Agrawal and Srikant [7]	Association rules; Apriori	Support and confidence formulation; level-wise mining with pruning by the anti-monotone property of support	Counts every record equally under one threshold; can need many passes and large candidate sets on dense data [8]
Han et al. [9]	FP-Growth	Mining from a compressed prefix tree without candidate generation	Single global support; the tree can consume considerable memory; no class or severity awareness
Pasquier et al. [12]; Wang et al. [13]	Closed itemsets; CLOSET+	Lossless condensed summary of frequent itemsets and their rules	Class label and record importance are not used
Liu et al. [5]; Tao et al. [14]	Multiple minimum supports; weighted support	Item-wise thresholds; weights on items with a usable pruning property	Thresholds and weights attach to items, not to outcome classes or to the severity of a record
Liu et al. [15]; Yin and Han [16]	CBA; CPAR	Rules that end in a class label, used to build classifiers	Same thresholds for every class; aimed at an accurate classifier
Webb [18]; Benjamini and Hochberg [19]	Significance testing; false discovery rate	Control of false discoveries when many patterns are tested	General tools; not built into the standard miners compared here
Geurts et al. [20]; Pande and Abdel-Aty [21]; Montella [22]; Kumar and Toshniwal [23]	Association rules, with clustering in [23], on crash data	Combinations of circumstances linked to hazardous locations and crash types	Off-the-shelf miner with a single global threshold and unweighted counts
Savolainen et al. [24]	Statistical injury-severity models	Review of methodological alternatives for severity analysis	Typically assume a functional form; do not directly yield simple combinations of circumstances

3.2 Research gap identified

The comparison points to a research gap that can be described in three parts. The first concerns rare severe outcomes. Apriori and FP-Growth count every record equally and apply one threshold to all items [7], [9]; multiple minimum supports attach thresholds to items [5] and weighted mining attaches weights to items [14], so neither derives the importance of a record from the seriousness of its outcome. The remedies for class imbalance are mostly resampling, cost-sensitive learning and algorithm-level changes designed for classifiers [6].

The second concerns the readability of the output. CBA and CPAR mine rules that end in a class label [15], [16], but they give every class the same thresholds, and closed itemsets reduce redundancy without using the label [13]. Standard algorithms return very many rules, many of them redundant or merely a consequence of a common

consequent, whereas practitioners need short and trustworthy lists [17]. The road-safety studies demonstrate that rules reveal combinations of circumstances that single-attribute tables cannot [20]–[22], but most of them apply an off-the-shelf miner with a single global threshold and unweighted counts, and statistical injury-severity models, though valuable, typically assume a functional form and do not directly yield simple combinations of circumstances [24].

The third concerns chance findings. Mining at low thresholds tests a very large number of candidate rules, and significance testing and false discovery rate control are available as general tools [18], [19], yet the standard miners compared here do not correct for the number of rules tested. Taken together, the algorithmic literature supplies the ingredients, but each was developed separately and none was designed for severity-labelled accident records, while road-safety work rarely modifies the mining algorithm itself. The research gap is therefore the absence of a single mining procedure that weights records by the seriousness of their outcome, lets a rare severe class have its own support threshold, restricts the search to rules that end in a severity class, removes redundant rules and protects against chance associations, while remaining efficient enough for large accident databases.

CONCLUSION AND FUTURE SCOPE

This paper reviewed the literature that leads from the foundations of association rule mining to its use on road-accident records, covering efficiency techniques, condensed representations, multiple supports and weights, class association rules, interestingness measures, class imbalance, statistical significance and accident-oriented data mining. A comparison of representative studies showed that the individual ingredients needed for severity-aware rule discovery exist, but that they have been studied separately, mostly on generic data, and that road-safety applications seldom adapt the mining algorithm to the rarity and unequal importance of severe crashes. The missing piece is an integrated procedure that is aware of severity, keeps the rules for rare severe outcomes, produces a short non-redundant rule list and guards against chance findings.

Future scope. Future work can address the gap in several ways. A pattern-growth miner could be extended so that every record is weighted by the severity of its outcome while the pruning property of support is preserved [9], [14], thresholds could be set for each outcome class instead of for each item [5], and the search could be limited to class-labelled rules as in classification-oriented miners [15]. The output could be shortened with closed itemsets [13] and interestingness measures [17] and then screened for chance associations with false discovery rate control [18], [19]. Learning the weights from crash-cost estimates and testing on data from other years and countries would strengthen any of these designs. One possible solution is a severity-weighted, class-rooted FP-Growth algorithm (SWC-FP) that combines class-specific supports, redundancy filtering and false-discovery-rate-controlled significance testing in a single framework.

REFERENCES

1. World Health Organization, Global Status Report on Road Safety 2018. Geneva: WHO, 2018.
2. Ministry of Road Transport and Highways, Government of India, Road Accidents in India 2019. New Delhi: Transport Research Wing, 2020.
3. U. Fayyad, G. Piatetsky-Shapiro, and P. Smyth, “From data mining to knowledge discovery in databases,” AI Magazine, vol. 17, no. 3, pp. 37–54, 1996.
4. R. Agrawal, T. Imielinski, and A. Swami, “Mining association rules between sets of items in large databases,” in Proc. ACM SIGMOD Int. Conf. Management of Data, 1993, pp. 207–216.
5. B. Liu, W. Hsu, and Y. Ma, “Mining association rules with multiple minimum supports,” in Proc. 5th ACM SIGKDD Int. Conf. Knowledge Discovery and Data Mining, 1999, pp. 337–341.
6. H. He and E. A. Garcia, “Learning from imbalanced data,” IEEE Transactions on Knowledge and Data Engineering, vol. 21, no. 9, pp. 1263–1284, 2009.
7. R. Agrawal and R. Srikant, “Fast algorithms for mining association rules,” in Proc. 20th Int. Conf. Very Large Data Bases (VLDB), 1994, pp. 487–499.
8. J. Hipp, U. Güntzer, and G. Nakhaeizadeh, “Algorithms for association rule mining – a general survey and comparison,” ACM SIGKDD Explorations Newsletter, vol. 2, no. 1, pp. 58–64, 2000.
9. J. Han, J. Pei, and Y. Yin, “Mining frequent patterns without candidate generation,” in Proc. ACM SIGMOD Int. Conf. Management of Data, 2000, pp. 1–12.
10. P. Fournier-Viger, J. C.-W. Lin, B. Vo, T. T. Chi, J. Zhang, and H. B. Le, “A survey of itemset mining,” WIREs Data Mining and Knowledge Discovery, vol. 7, no. 4, e1207, 2017.

11. M. J. Zaki and K. Gouda, “Fast vertical mining using diffsets,” in Proc. 9th ACM SIGKDD Int. Conf. Knowledge Discovery and Data Mining, 2003, pp. 326–335.
12. N. Pasquier, Y. Bastide, R. Taouil, and L. Lakhal, “Discovering frequent closed itemsets for association rules,” in Proc. 7th Int. Conf. Database Theory (ICDT), 1999, pp. 398–416.
13. J. Wang, J. Han, and J. Pei, “CLOSET+: Searching for the best strategies for mining frequent closed itemsets,” in Proc. 9th ACM SIGKDD Int. Conf. Knowledge Discovery and Data Mining, 2003, pp. 236–245.
14. F. Tao, F. Murtagh, and M. Farid, “Weighted association rule mining using weighted support and significance framework,” in Proc. 9th ACM SIGKDD Int. Conf. Knowledge Discovery and Data Mining, 2003, pp. 661–666.
15. B. Liu, W. Hsu, and Y. Ma, “Integrating classification and association rule mining,” in Proc. 4th Int. Conf. Knowledge Discovery and Data Mining (KDD), 1998, pp. 80–86.
16. X. Yin and J. Han, “CPAR: Classification based on predictive association rules,” in Proc. SIAM Int. Conf. Data Mining (SDM), 2003, pp. 331–335.
17. L. Geng and H. J. Hamilton, “Interestingness measures for data mining: A survey,” ACM Computing Surveys, vol. 38, no. 3, Art. 9, 2006.
18. G. I. Webb, “Discovering significant patterns,” Machine Learning, vol. 68, no. 1, pp. 1–33, 2007.
19. Y. Benjamini and Y. Hochberg, “Controlling the false discovery rate: A practical and powerful approach to multiple testing,” Journal of the Royal Statistical Society: Series B, vol. 57, no. 1, pp. 289–300, 1995.
20. K. Geurts, G. Wets, T. Brijs, and K. Vanhoof, “Profiling of high-frequency accident locations by use of association rules,” Transportation Research Record, no. 1840, pp. 123–130, 2003.
21. A. Pande and M. Abdel-Aty, “Market basket analysis of crash data from large jurisdictions and its potential as a decision support tool,” Safety Science, vol. 47, no. 1, pp. 145–154, 2009.
22. A. Montella, “Identifying crash contributory factors at urban roundabouts and using association rules to explore their relationships to different crash types,” Accident Analysis & Prevention, vol. 43, no. 4, pp. 1451–1463, 2011.
23. S. Kumar and D. Toshniwal, “A data mining approach to characterize road accident locations,” Journal of Modern Transportation, vol. 24, no. 1, pp. 62–72, 2016.
24. P. Savolainen, F. Mannering, D. Lord, and M. Quddus, “The statistical analysis of highway crash-injury severities: A review and assessment of methodological alternatives,” Accident Analysis & Prevention, vol. 43, no. 5, pp. 1666–1676, 2011.