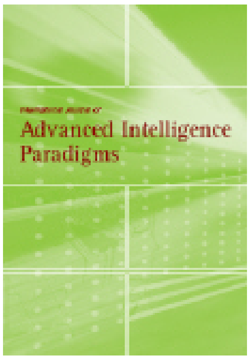
INDERSCIENCE PUBLISHERS
Linking academia, business and industry through research



International Journal of Advanced Intelligence Paradigms

ISSN online: 1755-0394 - ISSN print: 1755-0386
https://www.inderscience.com/ijaip

Map Reduce approach for road accident data analysis using data mining techniques

S. Nagendra Babu, J. Jebamalar Tamilselvi

DOI: 10.1504/IJAIP.2024.10062354

Article History:

Received:	24 March 2018
Last revised:	16 April 2018
Accepted:	16 April 2018
Published online:	22 February 2024

Map Reduce approach for road accident data analysis using data mining techniques

S. Nagendra Babu*

R&D Center,
Bharathiar University,
Coimbatore, India
Email: nagendra.csbu@gmail.com
*Corresponding author

J. Jebamalar Tamilselvi

Jaya Engineering College,
CTH Road, Prakash Nagar,
Thiruninravur, Chennai,
Tamil Nadu 602024, India
Email: jjebmalar@gmail.com

Abstract: Nowadays, the most life-threatening risk to humans is road accidents. Traffic accidents that cause a lot of damages are occurring all over the place. The best answer for these sorts of accidents is to foresee future accidents ahead of time, giving driver's odds to maintain a strategic distance from the perils or decrease the harm by reacting rapidly. The motivation behind this manuscript is to fabricate an anticipating structure that can resolve every one of these issues. This paper proposed hybrid N-clustering algorithm for performing clustering on road accident data and then improved association rule mining algorithm (IARM) for designing of several association rules for accident prediction and congestion control using machine framework (CCMF) and traffic congestion analyser using Map Reduce for efficient prediction of road accident on several factors using Map Reduce methods. To enhance the foreseeing precision, amended information is arranged into a few gatherings, to which characterisation investigation is connected.

Keywords: road accident prediction; Map Reduce; clustering; pre-processing; association rules; dataset.

Reference to this paper should be made as follows: Babu, S.N. and Tamilselvi, J.J. (2024) 'Map Reduce approach for road accident data analysis using data mining techniques', *Int. J. Advanced Intelligence Paradigms*, Vol. 27, No. 1, pp.1–17.

Biographical notes: S. Nagendra Babu is a research scholar in Bharathiar University, Coimbatore Tamil Nadu. He is doing his research in accident data analysis using data mining techniques. He received his Bachelor's degree from Sri Venkateswara University, Tirupathi, and his Master's degree in Computer Science from Sri Venkateswara University, Tirupathi, He possesses an excellent record all his academic performances. His areas of research are data warehousing and data mining. He has good number of publication in various Scopus indexed journals and free Scopus indexed journals.

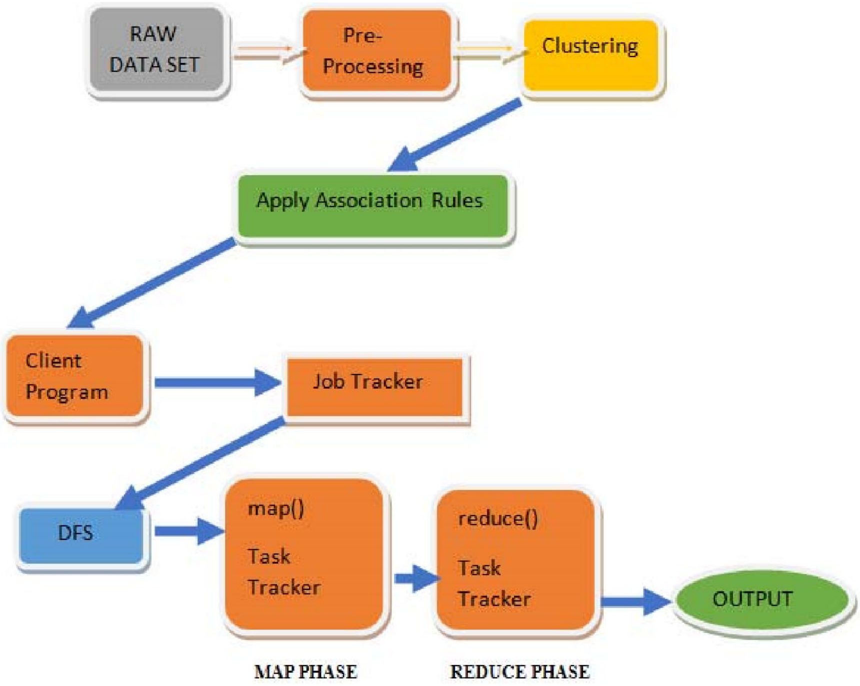
J. Jebamalar Tamilselvi received her PhD in 2009 from the Department of Computer Applications at Karunya University, Coimbatore, India. She received her BSc in Computer Science from Manonmanium Sundaranar University of Tamil Nadu, India in 2003 and MCA from Anna University, Coimbatore, Tamil Nadu, India in 2006. Her areas of interest include data cleansing approaches, data extraction, data integration, data warehousing and data mining. Her research has been accepted and published in 17 international journals, and 12 national and international conferences. She had been awarded the P.K. Das Memorial Best Faculty Award in 2014 by the Nehru Group of Institutions, Coimbatore and the Education and Research Award in 2015 by the Karunya University, Coimbatore.

1 Introduction

In current society, regular daily existence is personally worried about transportation making loads of issues on it. Among the issues, a standout amongst the most critical issues is about road accidents, and it is imperative to maintain a strategic distance from those accidents or diminish harm from them. Foreseeing conceivable road accidents can be an answer for those objectives. To foresee road accidents, we can utilise video picture from cameras out and about (Park and Ha, 2014), or different movement information to examine (Conejero et al., 2013). We are worried about investigating activity information with the goal that we can anticipate conceivable road accidents. In the proposed method totally three algorithms are developed in which the first algorithm N-centroid algorithm which is used as clustering algorithm which performs better clustering when compared to K-means algorithm. In association rule mining (ARM) a new enhanced association rule mining is proposed which is used to define association rules. The next two algorithms are for machine framework traffic data processing and Map Reduce-based algorithm for traffic data analysis. We mean to take road of these two issues and do proficiently forecast handling utilising the Map Reduce calculation. Along these lines, this paper presents preparing ventures of a parallel characterisation in light of Map Reduce and demonstrates the legitimacy of these means.

Here, in Figure 1, Map Reduce work flow of examining the information in parallel way it gives better execution. Here beneath we see the dispersion of the movement information into HDFS. On the off chance that a machine falls flat the information exhibit on that machine will be inaccessible for additionally preparing. So it is essential to build up an intense apparatus which can't be confined to capacity and preparing power abilities. Apache Hadoop (Conejero et al., 2013) is a dispersed system written in java to store immense measure of information and process the same. Today the majority of the organisations are utilising Hadoop innovation like Yahoo, Last.fm, Facebook, IBM and so on. Bigger bunches are accessible to clients to run their activity [Amazon (Zeng and McHugh, 2015), Grid5000 (Ashwini and Sunil, 2015)] with Hadoop on demand (HOD).

Figure 1 Process of proposed model (see online version for colours)



2 Literature survey

These segment investigations the craftsmanship in reflex strategies for creation stream rental number and evaluations the legitimate solicitations of configuration credit in rush hour gridlock modern. Example credit and assembling strategies have stayed utilised as a part of numerous parts of movement corporate. Grouping strategies are utilised to divider movement stream information into free present and stuck stream. The gathering method result is utilised to make disintegration hysterics and to grow a stream inhabitancy chart. Park and Ha (2014) propose a technique for making stream inhabitancy outline. In the initial step, dissemination information is isolated to free stream and packed stream in light of highlights of varieties in rush hour gridlock information time arrangement. Next, a base code technique is utilised to characterise symmetry states followed by the claim of a blended number advancement strategy to make piecewise lined stream inhabitancy fits. The principle significance of the advancement method is to get fits with least aggregate capriciousness.

Conejero et al. (2013) apply gathering strategies in the appearing of multi-administration speed-thickness relations. Grouping systems are utilised to perceive the disappointment focuses in a speed-thickness chart, speed-thickness information is then isolated in view of the perceived partitions, and wrinkled return strategies is utilised to make multi-administration speed-thickness relations.

Maitrey and Jha (2015) traffic information is divided to six groups, and the yields are contrasted and Highway Capacity Manual edges for level of administration. The continuous handling capacities of enormous information can precisely test road accidents, its prognostic capacity can adequately anticipate the event of activity episode, utilising microwave discovery frameworks, video observation frameworks, versatile recognition framework, we can manufacture a powerful security model to enhance the wellbeing of vehicles. At the point when security occurrence honed, and crisis save required, because of its comprehensive preparing and basic leadership capacity, speedy answer aptitude, huge information can incredibly recoup the ability of crisis protect, and lessen setbacks and property misfortunes.

In existing method a user has to manually perform pre-processing techniques on the raw dataset and they have to develop the trained dataset. From the trained dataset the user have to apply clustering technique for dividing the data into different clusters on type of dataset. Then based on the data user has to manually compare the dataset parameters in each case and has to record them separately.

After maintaining the data separately the data taken for accident prediction is manually checked with the trained dataset for accident prediction and for suggestions to the drivers. In this process as the entire process is done manually there are a lot of possibilities for false prediction based on errors in comparing trained dataset, mapping the types of accidents, etc.

3 Proposed method

3.1 Dataset

The dataset is considered from https://data.gov.in/dataset-group-name/road-accidents which provides a huge amount of data related to road accidents of different states in India and on this dataset pre-processing is applied for removing or filling of missed data. In this proposed method 121868 records of data for year 2017 is gone through pre-processing. In the proposed approach instead of manually performing the tasks the user after performing clustering operations the dataset is provided to machine which in turns compares with the dataset and prediction is accurate as all the association rules are applied by the machine to the dataset.

3.2 Pre-processing methodology

The arrangement of procedures utilised before the use of a big data mining strategy is named as information pre-processing for big data mining and it is known to be a standout amongst the most important issues inside the celebrated knowledge discovery from big data process. Since information will probably be flawed, containing irregularities and redundancies is not specifically pertinent for a beginning an information mining process. We should likewise specify the quickly developing of information age rates and their size in business, mechanical, scholastic and science applications. The greater measures of information gathered require more complex components to break down it. Information pre-processing can adjust the information to the prerequisites postured by every datum mining calculation, empowering to process information that would be unfeasible something else.

Table 1 Dataset used for road accident prediction

States	*2003*	*2004*	*2005*	*2006*	*2007*	*2008*	*2009*	*2010*	*2011*	*2012*	*2013*	*2014*	*2015*	*2016*	*2017*
Andhra Pr	9,800	10,981	10,026	10,953	9,979	10,634	9,887	10,681	11,211	8,353	9,621	5,644	5,351	5,639	5,852
Arunachal	144	84	111	82	88	115	114	128	0	0	0	0	0	75	288
Assam	784	746	947	946	1,010	849	1,034	1,345	1,552	1,553	1,817	1,777	1,898	1,999	2,212
Bihar	1,493	1,195	1,076	1,702	1,971	1,499	2,719	2,837	3,177	2,833	2,837	2,535	3,255	2,434	2,647
Chhattisg<	2,593	2,837	3,543	3,356	3,265	3,814	3,564	3,363	3,156	3,654	3,804	3,758	3,898	3,164	3,377
Goa	356	323	400	421	536	610	787	925	675	654	422	434	451	356	569
Gujarat	10,242	10,229	10,133	9,071	9,630	10,167	9,210	9,177	9,252	8,188	7,731	7,437	6,739	6,309	6,522
Haryana	2,529	2,931	2,917	3,202	3,752	3,611	3,693	3,436	3,425	3,108	3,253	3,247	3,389	2,887	3,100
Himachal	699	816	787	792	845	597	806	703	742	533	643	698	634	820	1,033
Jammu and Kashmir	956	1,087	1,060	1,286	745	971	1,054	930	892	948	922	767	837	537	750
Jharkhanc	1,146	1,476	1,580	1,438	1,766	968	1,165	1,077	1,332	1,358	1,758	1,252	1,480	1,704	1,917
Karnataka	8,269	7,972	8,062	10,004	15,034	15,723	12,500	13,215	12,522	13,012	11,898	13,308	10,254	11,462	11,675
Kerala	5,334	5,184	7,669	5,444	7,215	6,452	6,637	6,537	6,401	6,721	6,605	6,140	6,888	7,135	7,348
Madhya Pradesh	7,576	8,077	9,454	12,115	10,645	9,875	10,987	12,939	13,153	12,304	13,876	14,267	13,166	14,456	14,669
Maharash	11,618	13,853	11,831	11,957	13,402	13,307	12,230	12,767	13,149	12,846	12,029	11,760	11,184	9,052	9,265
Manipur	99	128	100	111	137	173	165	136	139	161	135	167	136	117	330
Meghalay	91	101	105	109	79	79	122	82	81	86	80	91	117	102	315
Mizoram	32	13	9	26	36	31	18	29	25	16	40	51	34	27	240
Nagaland	25	35	77	54	58	19	18	14	13	19	41	94	11	0	213
Odisha	1,324	1,466	1,466	2,088	2,198	1,964	2,386	2,062	2,129	2,333	3,433	3,507	4,074	3,328	3,541
Punjab	1,295	1,398	1,147	1,434	1,047	1,497	1,431	1,376	1,962	2,064	2,122	1,519	1,965	2,101	2,314
Rajasthan	2,430	2,596	2,380	2,175	2,870	2,581	2,913	3,119	2,625	2,723	3,029	3,774	3,638	3,695	3,908
Sikkim	39	66	101	34	26	36	159	49	170	32	83	71	85	70	283

3.2.1 Missing esteems ascription

One major suspicion made by information mining methods is that the informational collection is finished. The nearness of missing esteems is, notwithstanding, extremely basic in the procurement forms. A missing quality is a datum that has not been put away or accumulated because of a broken testing process, cost confinements or impediments in the securing procedure. Missing esteems cannot be stayed away from in information examination, and they have a tendency to make serious challenges for professionals.

Missing esteems treatment is troublesome. Improperly taking care of the missing esteems will effectively prompt poor learning extricated and furthermore wrong conclusions. Missing esteems have been accounted for to cause loss of productivity in the information extraction process, solid predispositions if the missed presentation instrument is misused and extreme confusions in information taking care of.

3.2.2 Clustering

For the most part, big data grouping strategies can be characterised into two classes (Chung et al., 2013): single machine grouping strategies what is more, various machine grouping systems, as of late the last mentioned draws more consideration since they are speedier and more adjust to the new difficulties of big data, single-machine strategies and grouping various machines incorporate distinctive systems.

3.2.3 Single-machine grouping

1 Data mining grouping calculations: the unsupervised order (grouping) is a basic data mining device for the examination of big data, which plans to merge the noteworthy class information objects with the goal that articles assembled in a similar bunch are comparable and steady as indicated by particular parameters.

It is hard to apply information mining grouping procedures in enormous data on account of the new difficulties. So with the considerable mass of information gave by the big data and the many-sided quality of bunching calculations (Potnis et al., 2014) which have high treatment costs, the inquiry that emerges is the manner by which to manage this issue and the most effective method to send bunching strategies big data to get comes about in a sensible time.

3.2.4 Multi-machine grouping

1 Parallel grouping: The preparing of a lot of information forces a parallel processing to accomplish brings about sensible time. In this area, we look at some parallel calculations and disseminated grouping used to treat big data, the parallel characterisation separates the information parcels that will be disseminated on various machines. This makes a person order to accelerate the figuring and increments adaptability.

2 Map Reduce-based grouping: Map Reduce is an errand dividing instrument (with huge volumes of information) for an appropriated execution on a substantial number of servers. Standard is to decay an errand into littler assignments. The assignments are then dispatched to various servers, and the outcomes are gathered and solidified.

An aggregate of 18% of the accidents happened amid high movement, 76% of accidents happened amid low activity, and 8% of accidents happened amid medium activity. A sum of 36% of accidents happened amid morning time, 17.5% of accidents happened toward the evening, 19.2% of accidents happened at night, and 33.4% of accidents happened amid evening time. 3.6% of accidents jumped out at the age gather youngsters, 73.4% of accidents struck adolescents, 25.6% of accidents jumped out at youth, 37.3% of accidents jumped out at moderately aged individuals, 23.5% of accidents jumped out at senior residents, and 27% of accidents age esteem is absent.

Table 2 Cluster size and its description

Cluster 1	*Cluster description*	*Count*	*Size (%)*
1	Two wheeler accidents on road intersections and curves near colonies and markets	3,181	27.48
2	Two wheeler accident occurred on highways near hill, fastest and agriculture land area	1,772	15.31
3	All fall height accidents with two or more injuries	1,928	16.66
4	Multiple vehicle accidents and fixed object hit accidents in no light condition	1,394	1204
5	Pedestrian hit cases	1,746	15.08
6	Vehicle roll-over accidents	1,553	13.42

Given a dataset DS, the relation between two different data clusters of different road accident cases M and N, where M and N are described by N categorical vehicles, can be computed as

$$d(M, N) = \sum\left(\delta\left(M_i, N_i\right)\right) \text{ for } i = 1 \text{ to } n$$

where

$$\delta\left(M_i, N_i\right) = 0, M_i = N_i \text{ and } 1, M_i \neq N_i$$

The below algorithm is used for clustering the road accident dataset which performance is better when compared to K-means algorithm.

Hybrid N-clustering algorithm ()

```
{
Input:
D = do1, do2, ..., don / set of n data objects. Then apply DBSCAN initially on N-clusters
Output:
A set of N number of clusters
Step-1: take a dataset as input
DBSCAN(DB, dist, eps, minPts) { C = 0
For each point P in database DB {if label(P) ≠ undefined then continue
Neighbours N = RangeQuery(DB, dist, P, eps)
if |N| < minPts then {label(P) = Noise and continue}
    C = C + 1
label(P) = C
    Seed set S = N \ {P}
Neighbours N = RangeQuery(DB, dist, Q, eps)
if |N| minPts then {
    $S = S \cup N$
    }   }   }   }
Step-2: apply N-centriod algorithm
    Pick N-initial centriods based on the distances divide the sorted data points into N number of
    equal partitions.
    Recalculate the centre of each cluster based on the data in the cluster.
    Repeat line 6 and line 7 until convergence.
    When the new cluster centres are the same as the clusters obtained in previous iteration,
    output the clustering results;
}
```

Table 3 Performance comparison

Dataset size	*Performance of clustering algorithm*	
	k-means algorithm	*Hybrid N-clustering algorithm*
1,000	1.1134	1.492482
2,000	1.3231	2.119413
3,000	2.1311	2.744253
4,000	2.3812	3.666876
5,000	2.1265	3.845595

3.3 Association rule setting

The ARM in information mining is a mainstream approach that is utilised to break down the offered dataset to find fascinating examples or connections between the different things in the dataset (Bhagattjee, 2014). The ARM strategy produces an arrangement of association rules winning between the different things of the given dataset in view of the quantity of events of these things blend in the dataset.

Here a new improved algorithm is defined for ARM for road accident data prediction. The proposed IARM algorithm is much improved in frequent item set mining when compared to FP-growth algorithm. This algorithm defines how rules can be set to define new strategies for mining frequent item patterns.

```
//Generate vehicle ids, accident types, their accident IDs
Algorithm IARM()
{
V1 = find_frequent_1_itemsets(P);
For (k = 2; Lk – 1 ≠ Φ; k++)
{
Ck = candidates generated from Lk – 1;
x = Get_item_min_sup(Ck, V1);
Tgt = get_accident_ID(x);
For each accident t in Tgt Do
Increment the count of all vehicles in Ck that are found in Tgt;
Lk = vehicles in Ck ≥ defined_Tgt;
End;
}
```

An association control is utilised to characterise the connection between any two things in the given dataset. Think about three things P, Q and R. The connection $\{P, Q\} \rightarrow R$ say that if a man purchases two things P and Q together, at that point he/she will in all likelihood purchase the thing R moreover. That is, the relations between the things are produced by recognising the different examples inside the dataset. The ARM procedure (Manikandan and Ravi, 2014) comprises of two phases as takes after:

1 Identify the item set that happen regularly in the dataset – the successive item set are those that have a help esteem (sup(item)) equivalent to or more prominent than the base help esteem that is pre-characterised (Zeng, 2015). The help estimation of item set is figured as the quantity of exchanges that contains that thing. In the above illustration support of $\{P, Q\}$ is figured as what number of exchanges have both P and Q.

2 Association manage age utilising regular item set: In this stage the fascinating principles are produced by computing the certainty factor for all the regular item set that are created in past stage. The certainty esteem for the above case govern of $\{P, Q\} \rightarrow R$ will be $\text{sup}(\{P, Q\}) / \text{sup}(R)$.

3 Given an informational collection D of n vehicles where every vehicle $T \in D$. Let $I = \{I1, I2, \ldots, In\}$ is accident data. A accident set A will happen in T if and just if $A \subseteq T$. AUB is and affiliation govern, gave that $A \subset I$, $B \subset I$ and $A \cap B = \varnothing$. In association manage mining, support, certainty and lift measurements are notable fascinating measure which is utilised as a part of picking solid affiliation rules.

For instance consider different data clusters on road accident data then required association rules are determined as

{Road_type = "parkway" AND Road_feature = "crossing point" AND
Area_around = "colony"} and {Victim_injured = "1"}

{Road_type = "nearby" AND Road_feature = "crossing point" AND
Area_around = "market"} and {Victim_injured = " > 2"}

{Road_type = "nearby" AND Road_feature = "crossing point" AND
Area_around = "market"} and {Victim_injured = " > 2"}

{Road_type = "highway" AND Area_around = "hill" AND Road_feature = "bend"
AND Victim_injured = " > 2"} and {Accident_severity = "critical"}

$$\{\text{Road\_type} = \text{"local" AND Area\_around} = \text{"market" AND}$$
$$\text{Victim\_injured} = \text{"1"}\} \text{ and } \{\text{Age} = \text{"Young"}\}$$

$$\{\text{Road\_type} = \text{"local" AND Area\_around} = \text{"colony" AND}$$
$$\text{Victim\_injured} = \text{"2"}\} \text{ and } \{\text{Age} = \text{"child"}\}$$

The performance of the proposed ARM algorithm is compared with FP-growth algorithm. The performance levels are compared in seconds.

Table 4 Performance levels in time

Type	*FP-growth*	*IARM*
DATASET < 1,000 records	3.66s	3.03 s
DATASET < 5,000 records	8.87 s	3.25 s
DATASET < 10,000 records	34 m	5.07 s
DATASET < 2 lackds	4+ hours (never finished, crashed)	8.82 s

3.3.1 Map Reduce approach for ARM

The association rules and the age of guidelines are generally utilised and they confront numerous issues and the significant one is the accessibility of substantial information and multi-dimensional datasets (Zeng and McHugh, 2015). A Map Reduce work as a rule parts the info information into different pieces and each of these are handled by the guide errands in parallel way. The Mapper maps the little undertakings by influencing utilisation of the key and incentive to combine idea and the yields are arranged. At that point the reducer lessens the acquired yields from the maps to get the last yield (Ashwini and Sunil, 2015). The Map Reduce structure contains a single job tracker as the ace and a solitary task tracker as the slave for each bunch hub. All info and yield in Map Reduce are <key, value> sets. The Hadoop is a Java-based appropriated programming condition supported by Apache that can be utilised to process and handle a lot of information. Hadoop has been made utilising the idea of Map Reduce for extensive preparing by utilising an extensive number of hubs and bunches.

3.4 Applying Map Reduce methods

In the proposed method initially we take the dataset and we give that dataset for pre-processing, after pre-processing we will get a clean dataset which contains heterogeneous data types. In the manuscript two algorithms are proposed congestion control using machine framework (CCMF) which is used for considering traffic data by the machine framework. Another algorithm traffic congestion analyser (Pirttikangas and Riekki, 2015) using Map Reduce (TCAMR) which intern analyse the traffic data for prediction of road accidents. Take all those data and go for map-reduced-based unsupervised clustering, from there we will get an processed clusters, and then apply association rules then we are already designed a machine with all the possible inputs and also tested, after pre-processing (Sinnott, 2015) we will able to generate the prediction. In case if we may missed any kind of special input or situation we will again generate the new training and testing mechanisms to the machine. Like that we are continuing the procedure and we will able to overcome the problems associated from the traffic problems like accidents and other traffic problems.

Algorithm CCMF ()

```
{
    1.  Initially take the training data $t_1, t_2, …, t_n$
    2.  Give each and every aspect of training data to the machine
    3.  After the completion of training go for testing
    4.  Give test data to the machine
        a.  If any failure
        b.  Go back to step 1 and include the failure case to train data
        c.  Or test all the test case
    5.  And derive the prediction's
}
```

Algorithm TCAMP ()

```
{
Input traffic dataset
Output predicting the accidents
Step-1:  take the traffic dataset
Step-2:  apply pre-processing
Step-3:  apply Map function with an clustering algorithm
    1.  Partitioning the traffic data
    2.  Send the partitions onto different machines
    3.  Map the each partitions value into a key value pair
Step-4:  apply Reduce function
    1.  Shuffling
    2.  Reduce into unique key value pairs
    3.  Get the clusters
After
Step-5:  take the clusters and apply the regressions
}
```

4 Experimental setup

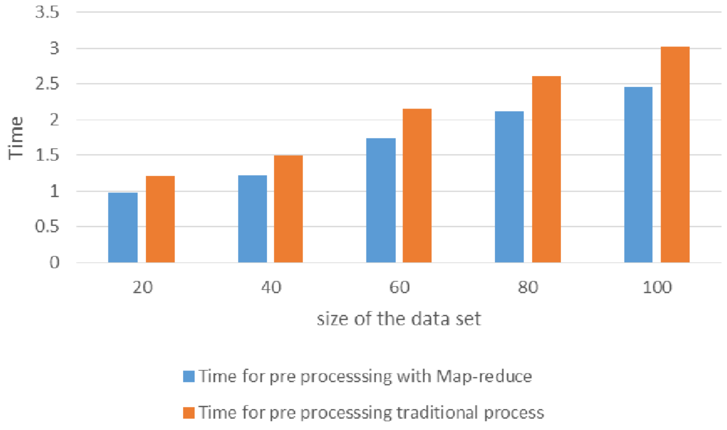
Here for experimental setup we use a machine with 16 GB Ram, 1TB HDD, with Ubuntu 16.04 lts and it was with Hadoop and apache mahout installed in it. And also the machine consists of two cores. The attributes considered in the dataset are listed in Table 5.

Table 5 Road accident analysis parameters

Attribute name	*Values*	*Description*
Accident ID	Integer	Identification of accident
Accident type	Fatal, injury, property damage	Accident type
Driver age	<20, [21–27], [28–60] >61	Driver age
Driver sex	M, F	Driver sex
Driver experience	< I, (2–4), >5	Driver experience
Vehicle_age	[l–2], [3–4], [5–7], >10	Service year of the vehicle
Vehicle_type	Car, trucks, motorcycles, other	Type of the vehicle
Light_condition	Daylight, public lighting, night	Light condition
Weather_condition	Normal weather, rain, fog, wind, snow	Weather conditions
Road_Condition	Highway, ice road, collapse road, unpaved road	Road conditions
Road age	[l–2], [3, 5], [6–10], [11–20] >20	The age of road
Time	[0–6], [6–12], [12–18], [18.00]	Accident time
Particular_area	School, market, temple	Where accident occurred in school or market areas
Season	Autumn, spring, summer, winter, rainy	Seasons of year
Accident_causes	Alcohol effects, fatigue. Losses of control, speed, pushed by another vehicle, brake failure	Causes of accident
Number of death	l, [2–5],(6–10), >10	Number of deaths

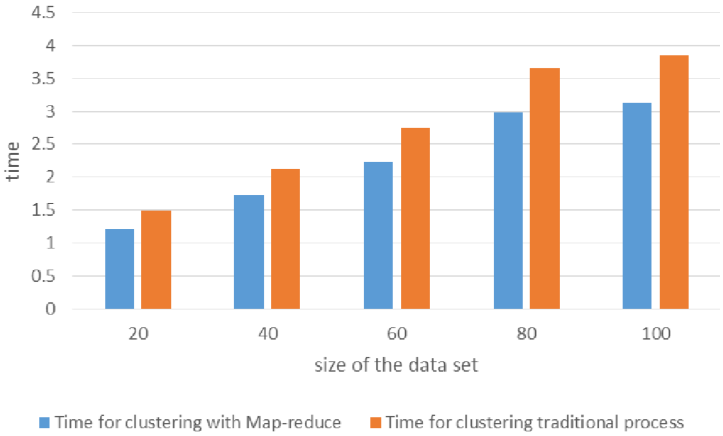
Here in Figure 2, time for pre-processing shows the comparative performance analysis for pre-processing stage with Map Reduce approach (Kumar and Toshniwal, 2015) and without Map Reduce obviously the parallel processing technique that is Map Reduce take less time compared to the traditional methods.

Figure 2 Pre-processing time (see online version for colours)



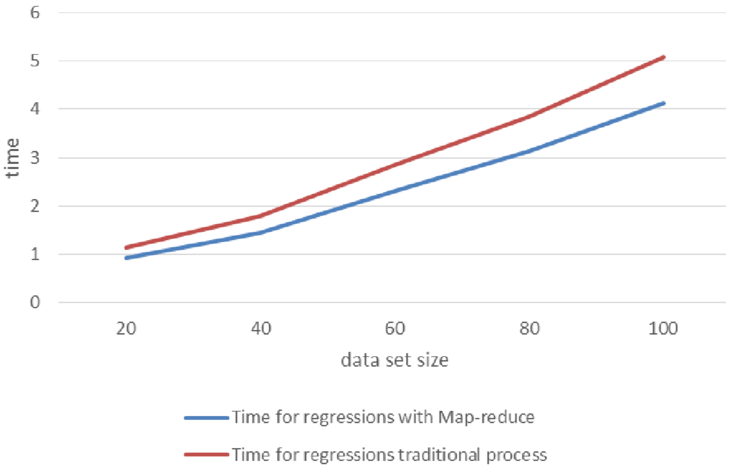
Here in Figure 3, time for clustering shows the comparative performance analysis for clustering stage with Map Reduce approach and with out map-reduce obviously the parallel processing technique that is Map Reduce take less time compared to the traditional methods.

Figure 3 Time for clustering (see online version for colours)



Here in Figure 4, time for regression shows the comparative performance analysis for regression stage with Map Reduce approach and with out Map Reduce obviously the parallel processing technique that is ma-reduce take less time compared to the traditional methods.

Figure 4 Time for regression (see online version for colours)



Here in Figure 5, accuracy comparison shows the accuracy of predicting of existing VDS method and the proposed technique the results shows that the when the size of dataset increase the accuracy of the proposed method increase.

Figure 5 Accuracy comparison (see online version for colours)

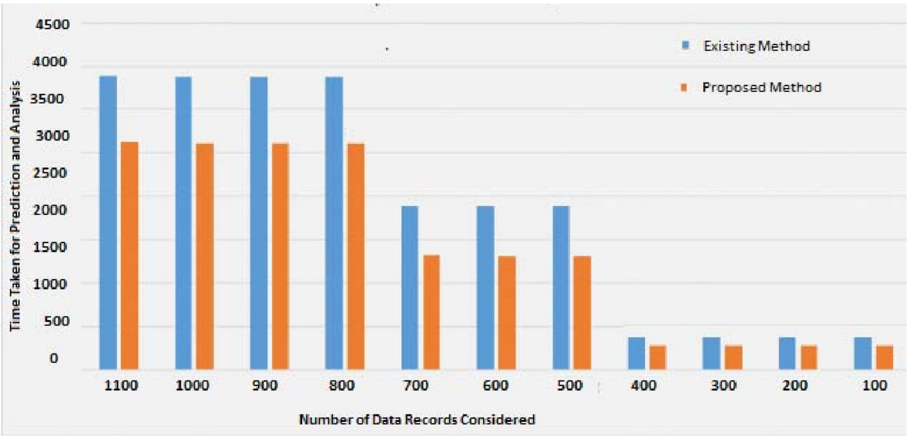
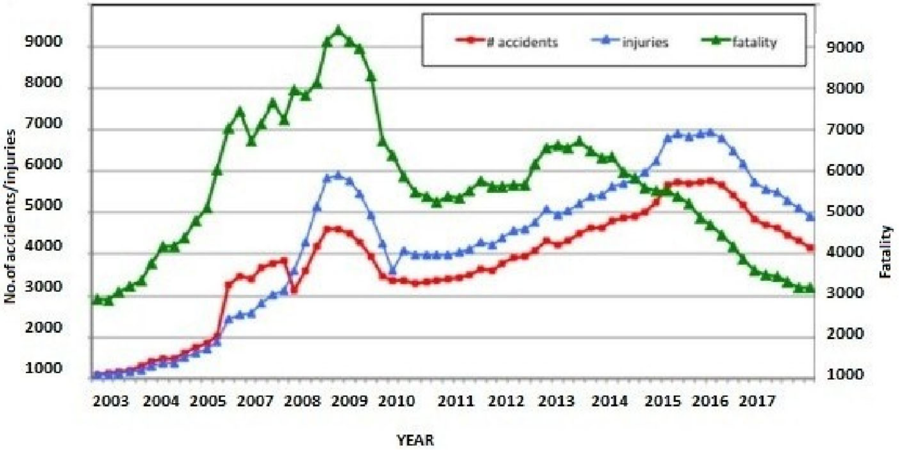


Figure 6 performs analysis on no. of accidents occurred from the previous data considered. The analysis shows the count of persons undergone accident/injury/fatality in a year. The proposed method clearly illustrates the data and gives predictions to the drivers for maintaining safety and to take care while travelling.

Figure 6 Analysis of no. of accidents/injuries/fatality per year (see online version for colours)



4.1 Analysis

Road accidents and wounds happen due to human blame or vehicle blame or framework blame or now and then mix of these variables. Every one of these components independently or in mix may cause mischance. It was seen from the dataset that accidents for the most part happened as a result of mix of human blame and vehicle blame as appeared in Table 6. Human alone factors, for example, 'head protector and safety belt not utilised' are not detailed in the FIRs and all things considered are not known. Table 6 presents top three contributing variables for accidents, most elevated being impulsive driving of the general population.

Table 6 Top factors for road accidents

Contributing factor	*Percentage of accidents (%)*
Rash driving	62.57
Object hit	26.67
Lane change	8.1

Figure 7 Accidents by speed limit (see online version for colours)

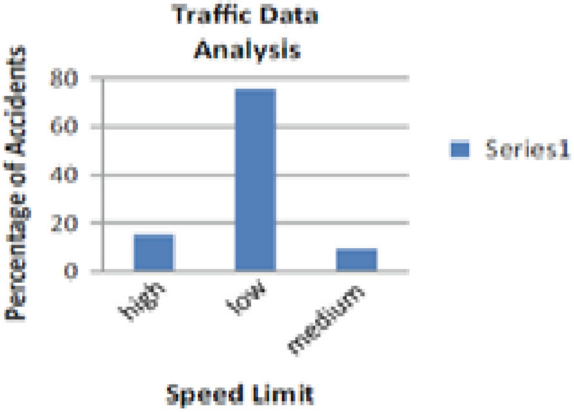
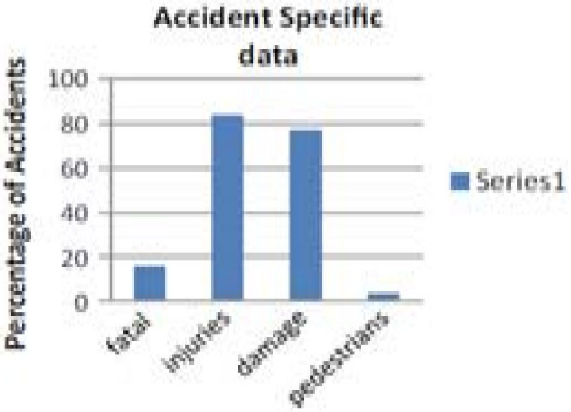
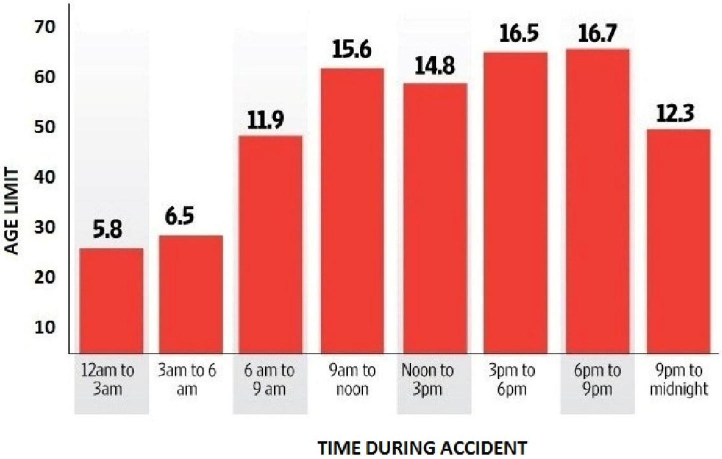


Figure 8 Accidents by injury severity (see online version for colours)



Based on Figure 9, we can propose a solution for road accident prediction based on time on the day and the age of the driver who is driving a vehicle which meets with accident. Based on the above graph users are guided about the time in which majority accidents takes place and also warn about age limit for driving.

Figure 9 Final road accident prediction based on time and age (see online version for colours)



5 Conclusions

This paper talks about the road accident investigation, which is unmistakably identified. In this paper, we proposed a system for examining accident designs for various kinds of accidents out and about which influences utilisation of Hybrid grouping and enhanced relationship to administer mining calculation. We played out a few tests on road accident information to gauge the speed and scale up of usage of algorithm traffic data analysis in Sparks' MLlib. We discovered much superior to expected outcomes for our trials. The outcomes exhibit that the proposed approach is profoundly versatile and could give significant data that can assist the coordination management with improving the exhibitions of transport quality and roads better improvement.

References

Ashwini, E.M. and Sunil, M.E. (2015) 'Tadoop: traffic analysis and traffic solution using Hadoop', *International Journal of Innovative Research in Computer and Communication Engineering*, October, Vol. 3, No. 7, pp.46–50.

Bhagattjee, B. (2014) *Emergence and Taxonomy of Big Data as a Service*, Submitted to the System Design and Management Program on 30 February 2014 in partial fulfilment of the requirements for the degree of Master of Science in Engineering and Management.

Chung, D., Rui, X., Min, D. and Yeo, H. (2013) 'Road traffic big data collision analysis processing framework', *2013 7th Int. Conf. Appl. Inf. Commun. Technol.*, October, pp.1–4.

Conejero, J., Burnap, P., Rana, O. and Morgan, J. (2013) 'Scaling archived social media data analysis using a Hadoop cloud', *IEEE.*

Kumar, S. and Toshniwal, D. (2015) 'A data mining framework to analyze road accident data', *Journal of Big Data*, 21 November, Vol. 2, No. 26, pp.1–8.

Maitrey, S. and Jha, C.K. (2015) 'Handling big data efficiently by using Map Reduce technique', *IEEE Int. Conf. Comput. Intell. Commun. Technology*, February, pp.703–708.

Park, S-h. and Ha, Y-g. (2014) 'Large imbalance data classification based on MapReduce for traffic accident predication', *IEEE International Conference on Innovative Mobile and Internet Services in Ubiquitous Computing*, pp.45–49.

Pirttikangas, S. and Riekki, J. (2015) 'Low latency analytics for streaming traffic data with Apache Spark', Maarala, A., Rautiainen, M. and Salmi, M. (Eds.): *Big Data* (*Big Data*), *2015 IEEE International Conference*, pp.2855–2858.

Potnis, A.A., Pandit, H.V. and Deshpande, S.S. (2014) 'Vehicular travel initiated sustainable USB mobile charging and travel analytics system', *Big Data and Cloud Computing* (*BdCloud*), *2014 IEEE Fourth International Conference*, pp.620–624.

Sinnott, R.O. (2015) 'Accident black spot identification and verification through social media', *2015 IEEE International Conference on Data Science and Data Intensive Systems*, epartment of Comput. & Inf. Syst., Univ. of Melbourne, Melbourne, VIC, Australia, Shuangchao Yin, Sydney, NSW, 11–13 December, pp.17–24.

Zeng, G. (2015) 'Application of big data in intelligent traffic system', *IOSR Journal of Computer Engineering*, January–February, Vol. 17, No. 1, pp.1–4.

Zeng, G. and McHugh, D. (2015) *Traffic Prediction and Analysis using a Big Data and Visualisation Approach*, Department of Computer Science, Institute of Technology Blanchardstown, 10 March.