

Trend Analysis Using Agglomerative Hierarchical Clustering Approach for Time Series Big Data



P. Subbulakshmi, S. Vimal, M. Kaliappan, Y. Harold Robinson, and Mucbeol Kim

1 Introduction

Road traffic accident (RTA) is an important factor to consider in research as it contains many factors and injuries lead to disability and other personal affections [1]. A report generated by World Health Organization shows that about 2 million accidents are happening throughout worldwide and it may lead to various segment of road safety and injuries [2]. The main focus on road accident is to ensure safety in road and emergency reporting [3]. The time sequence data analysis helps in encouraging the user to predict the emergency service, and accident rate can be predominantly reduced in there forth [4]. Data mining technique has been applied to statistically survey the accident rate [5]. The challenges of big data are capture, storage, search, sharing, transfer, analysis, and visualization [6]. These data are in various formats such as structured, unstructured, and semi-structured [7]. These data

P. Subbulakshmi
School of Computing, Scope, VIT University, Chennai, India

S. Vimal
Department of Computer Science and Engineering, Ramco Institute of Technology, Rajapalayam, Tamil Nadu, India

M. Kaliappan
Department of Computer Science and Engineering, Ramco Institute of Technology, Rajapalayam, Tamil Nadu, India

Y. H. Robinson
School of Information Technology and Engineering, Vellore Institute of Technology, Vellore, India

M. Kim (✉)
School of Computer Science & Engineering, Chung-Ang University, Dongjak-gu, Seoul, South Korea

are not handled by the traditional data processing tools [8]. In this paper, we take the huge amount of data generated in the road accidents for finding the factors causing them. This can be handled by the clustering and association rule mining techniques [9]. The trend analysis in PTS is more helpful for finding the factors of the accident since accident ratio increases every year. This can be efficiently identified with the time sequence data in every location. These data are needed to be normalized and then used for the processing with the help of the data mining techniques [10].

2 Related Work

Time series performs a group of related data points gathered within the specified interval [11]. Monthly basis road accident is measured to store the time series data for revealing the future trend. This will help to identify the dissimilar regions of road accidents for providing the trend analysis [12]. This analysis contains 11 leading cities in India for using the dataset. It is very hard to analyze the time series data of the leading cities independently and also relate to the nature and trend of the road accidents in the similar other cities in India [13].

3 Proposed Work

3.1 Data Preprocessing

Data preprocessing is a prior task is to analyze the data in every data handling techniques. The techniques of data preprocessing remove the unwanted noise or other constraints in the network. Here, this time sequence data was preprocessed, and an analysis has been normalized [14]. The data transformations have been performed to implement data available for the time series-related analysis. The time series information needs the enhanced data preprocessing in spite of getting the data useful for the specific analysis. The time series information uses the normalization technique for assisting the difficulties in the preprocessing technique. Hence, the analysis is used to get the efficient result for normalizing the time series information based on the time series.

3.2 Similarity Measure for Time Sequence

Dynamic time warping (DTW) measures the correlation between two time sequence data objects even if their lengths are not the same. DTW is used to minimize the metrics of two time sequences $r_i = \{r_1, r_2, \dots, r_n\}$ and $f_i = \{f_1, f_2, \dots, f_m\}$ which are of length x and y , respectively, and have to align r_i and t_j . The dynamic programming

establishes a cluster approach using the infinity matrix, and the parameters are computed in Eqs. (1) and (2):

$$\text{mean}(f_i) = 0 \quad (1)$$

$$\text{SD}(f_i) = 1 \quad (2)$$

where $\text{mean}(f_i)$ is the mean value and $\text{SD}(f_i)$ is the standard deviation for producing the normalization time series. The normalization is computed in Eq. (3).

$$\text{Nor}_{f_i} = \sum_{t=1}^n \frac{\text{time}_i - \text{mean}(f_i)}{\text{SD}(F_i)} \quad (3)$$

3.3 Hierarchy-Based Cluster Analysis

Cluster analysis is done using the similar objects to be grouped in a single forum or groups based on their attributes and properties. A lot of clustering algorithms such as agglomerative clustering are used in statistical data prediction. There may be a variety of algorithms for clustering a time sequence data. Agglomerative hierarchical clustering algorithm is designed to map reduce framework for clustering of time sequence data. The space and time complexity using agglomerative hierarchical clustering is $P(n_3)$, and the other one is $P(2n)$. The proposed AHCTB is used to predict the road accident data as shown in Fig. 1.

Algorithm for simple hierarchical agglomerative clustering

```

Input: Preprocessed Road Accident dataset
Output: PTS for each cluster
Initialization:
    Set line, month, district as Object
Preprocessing:
Tokenize dataset
Line=get line from the dataset
Month=get the Month value from the line
District=get the District value from the line
Agglomerative hierarchical clustering algorithm for mapper and
reducer: (mapper)
    Do until reach the Month & District
    Set district = district value of line
    If Month Exists
    Set Month as Key
    Set month = month value of line
    End If
    End
    Merge month and district value
    Set month and district value into another variable
    Map the variable to the reducer
    (Reducer)
    Set sum=0
    For each get value from the key
    Calculate Euclidean mean for each district = PTS

```

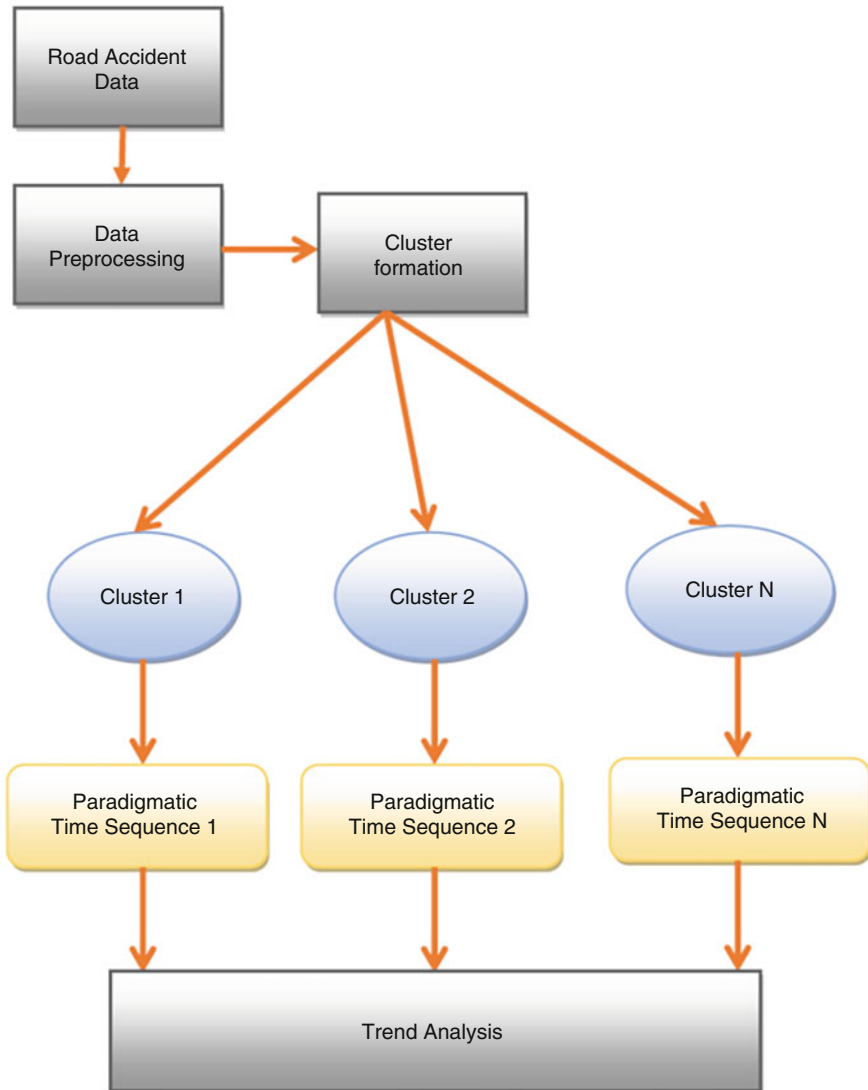


Fig. 1 Proposed system model

4 Performance Evaluation

4.1 Monthly Prediction

The clustering of districts makes the time sequence data to be visualized to perform the normal conditions over the trend analysis. In Fig. 2, the accident rate is high

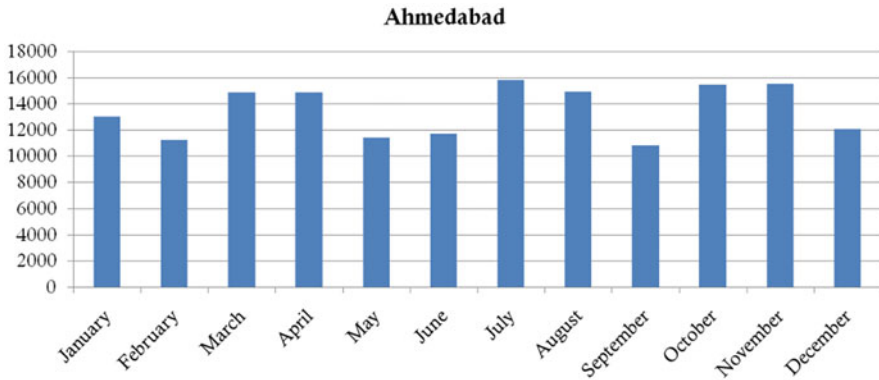


Fig. 2 Accident rates in Ahmedabad city

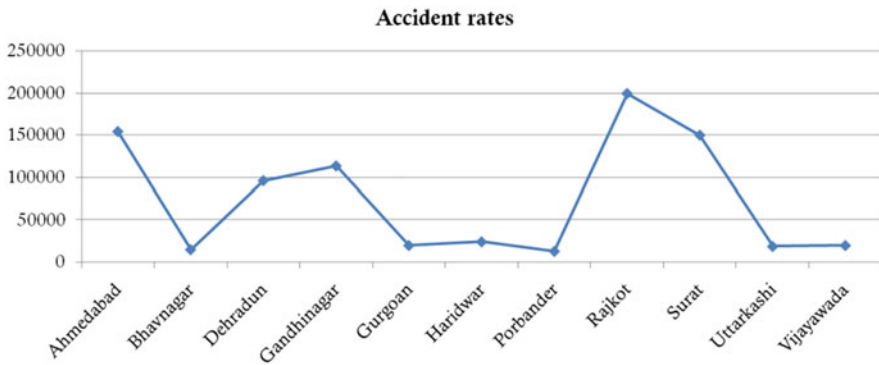


Fig. 3 Accident rates for Indian cities

because it is a metropolitan city which has a high population rate, results in high traffic which tends to maximum accident occurrence. Here, in each month, the trend rate varies depending on the accident count. The high peak of accident rate is recorded in the months of July and November.

Figure 3 demonstrates the accident rates for 1 single year for 11 important cities, and it illustrates that the cities of Ahmedabad, Surat, and Rajkot are having highest amount of accident rates. The cities of Bhavnagar, Gurgaon, Haridwar, Porbandar, Uttarkashi, and Vijayawada are having the lowest amount of accident rates. The cities of Dehradun and Gandhinagar are having the medium amount of accident rates.

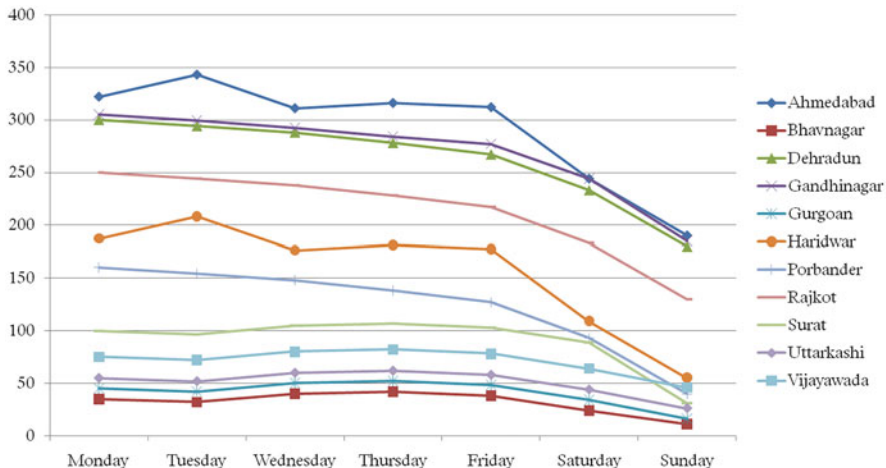


Fig. 4 Weekly analysis

4.2 Weekly Analysis

Fig. 4 demonstrates the weekly analysis for the road accident rates for the 11 Indian cities in detailed manner.

Figure 5 analyzes the time of accident percentage for the cities, and it is measured that at the time of 12–2 pm, the accident rates are high. At the time period of 2–4 pm, the accident rate is around 19 percentages. At the time between 8 and 10 am, the accident rate is 15%, and in all the other timings, the accident rate percentage is 10 or less.

5 Conclusion

The progress of the system includes preprocessing of heterogeneous data which are classified and clustered based on the district-wise grouping of the road accident rates. Consequently, in each cluster, attributes are merged based on the monthly analysis, and paradigmatic time sequence is predicted which is then fed for the trend analysis of the accident rate. Here, AHCTB is proposed for the clustering of each district, and trend analysis is done to each cluster using PTS. The trend analysis shows the variation of the accident rates in each cluster. It also shows that accidents are prone to increase in metropolitan and industrial cities where the population and transportation are very high in nature, whereas in rural areas, the accident rate is low.

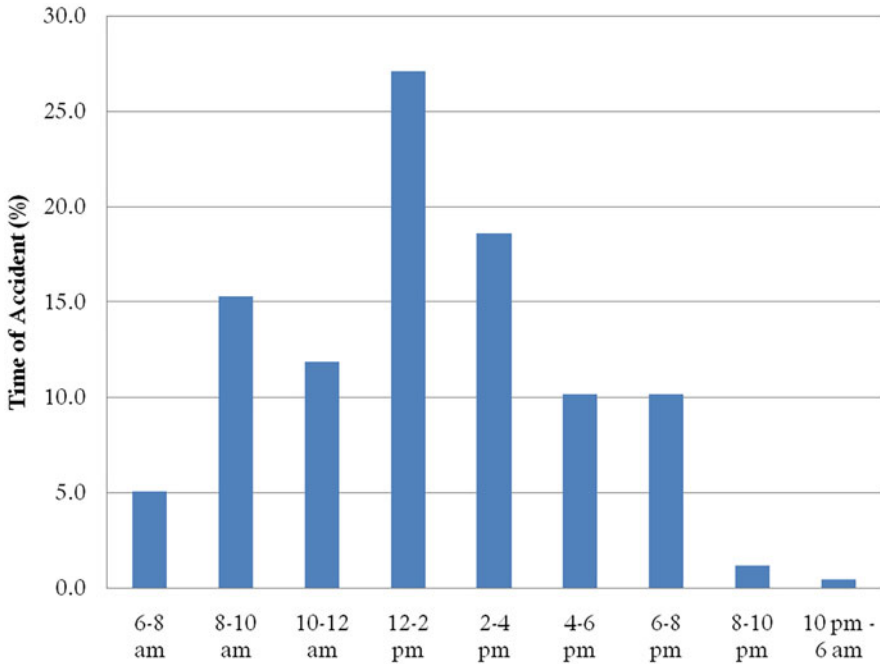


Fig. 5 Time of accident

Acknowledgement This work was supported by the National Research Foundation of Korea (NRF) grant funded by the Korean government (MSIT) (No. 2020R1F1A1076976).

References

1. J. Abellan, G. Lopez, J. Ona, Analysis of traffic accident severity using decision rules via decision trees. *Expert Syst. Appl.* **40**, 6047–6054 (2013)
2. S. Kumar, D. Toshniwal, Analyzing road accident data using association rule mining, in *International Conference on Computing, Communication and Security, ICCCS-2015*, vol. 20, pp. 30–40
3. S. Kumar, D. Toshniwal, A novel framework to analyze road accident time series data. *J. Big Data* **30**, 5004–5020 (2016)
4. L. Wang, H.-p. Lu, Y. Zheng, Z. Qian, Safety analysis for expressway based on Bayesian network: A case study in China. *IEEE Commun. Mag.* **18**(12C), 438–434 (2014)
5. J. de Ona, R.O. Mujalli, F.J. Calvo, Analysis of traffic accident injury severity on Spanish rural highway using Bayesian networks. *ScienceDirect Accid. Anal. Prev.* **43**, 402–411 (2011)
6. A. Pakgozar, R.S. Tabrizi, M. Khalili, A. Esmaili, The role of human factor in incidence and severity of road crashes based on the CART and LR regression: A data mining approach. *ScienceDirect Procedia Comput. Sci.* **3**, 764–769 (2014)
7. S. Kumar, D. Toshniwal, A data mining approach to characterize road accident locations. *J. Mod. Transp* **24**(1), 62–72 (2016)

8. Y. Lv, Y. Duan, W. Kang, Z. Li, F.-Y. Wang, Traffic flow prediction with big data: A deep learning approach. *IEEE Trans. Intell. Transp. Syst.* **16**, 2 (2015)
9. J. de Oña, G. López, R. Mujalli, F.J. Calvo, Analysis of traffic accidents on rural highways using Latent Class Clustering and Bayesian Networks. *ScienceDirect Accid. Anal. Prev.* **51**, 1–10 (2013)
10. S. Regine, C. Simon, A. Maurice, Processing traffic and road accident data in two case studied of road operation assessment. *ScienceDirect Transp. Res. Procedia* **6**, 90–100 (2015)
11. S. Kumar, D. Toshniwal, A data mining framework to analyze road accident data. *J. Big Data* **2**, 26 (2015)
12. D. Kee, G.T. Jun, P. Waterson, R. Haslam, A Systemic analysis of South Korea Sewol ferry accident – Striking a balance between learning and accountability. *Appl. Ergon.* **1**, 1–14 (2016)
13. L. Ramos, L. Silva, M.Y. Santos, J.M. Pires, Detection of road accidents accumulation zones with a visual analytics approach. *ScienceDirect Procedia Comput. Sci.* **64**, 969–976 (2015)
14. S. Kumar, D. Toshniwal, Analysis of hourly road accident counts using hierarchical clustering and cophenetic correlation coefficient (CPCC). *J. Big Data* **3**, 13 (2016). **30**, 20–56