An Innovative Context to Explorate Road Accident Time Series Data

Mekala.R¹, Sujitha.K², Susila Devi.V³, Thirukumaran.P P⁴

¹Assistant Professor, Department Of CSE, K.S.Rangasamy College of Technology, Namakkal, Tamil Nadu, India.
²³⁴Students, Department Of CSE, K.S.Rangasamy College of Technology, Namakkal, Tamil Nadu, India.

Abstract—Road accident data analysis is an important aspect in identifying the causes of accidents. The factors associated with the accident are identified and it helps to take preventive measures to avoid accidents. Using data mining techniques and traditional statistical techniques, the road accident data were analyzed. The road accident data from different regions were collected and various studies have been done. Since road accidents are uncertain, their impacts are unpredictable. The impacts differ for every road accident. Hence there is a need for road safety in all the regions where the road accident trend is increasing. The varying of road accidents can be studied using time series analysis. In this paper, we have suggested the analysis based on the time series data of road accidents in various districts of Tamil Nadu. This analysis divides the time series into different clusters. For each cluster, a Representative Time Series (RTS) is formed by time series merging algorithm. Multi view point clustering is done for the objects within the cluster. The consequence exposes that road accident trend is going to increase in certain clusters and those regions should have prime apprehension to take deterrent measures to evade accidents.

Keywords—Road accident; data mining; time series; multi view point similarity clustering

I. INTRODUCTION

Road traffic accident (RTA) is one of the significant apprehensions of research as it involves casualty, personal injuries that can lead to full or partial infirmity and property damage. A report by World Health Organization (WHO) reveals that there are 1.2 million incurable and about four times injured road accidents every year across the world. Road and traffic safety is a term associated with road accidents. The crucial focus of road safety is to deliver some preventive measures that can be supportive in reducing RTAs.

Road accident data exploration is a vital aspect that has been flourishes in recognizing different influences associated with road accidents. As soon as the associated road accident factors are identified the equivalent actions can be taken to overcome the accident rate and to apply some precautionary measures. Road accident data exploration is mainly based on two sorts: statistical methods and data mining methods. Numerous studies on road accident data analysis used traditional statistical methods and data mining methods. Data mining techniques have firm advantages over traditional statistical techniques. Data mining techniques do not entail certain conventions between dependent and independent variables which are required in traditional statistical techniques. Also, data mining techniques is accomplished with supervision large dimensional data whereas statistical techniques have some boundaries.

The road accidents may have different impact for different type of accidents at different locations. Also, road accidents are also varying district wise and it may happen that certain districts have similar nature of road accidents. Time series organize a series of data points composed or appraised at fixed interval. Monthly road accident counts of road accident for a certain period of time also constitute a time series data. The time series data of road accidents is very vital to study as it can expose the future trend of road accidents. This future trend can help in ascertaining the different regions where the road accidents is tend to increase or decrease so that
deterrent dealings can be taken. This study uses time series data of various districts of Tamil Nadu. It is difficult to scrutinize all the time series data of different districts individually and also based on the assumption that nature and trend of road accidents can be similar in some districts.

In previous work, authors strained to eradicate the heterogeneity in road accident data using data mining techniques and used a context using clustering and association rule mining techniques that is capable to remove the heterogeneity from road accident data. Those techniques can certainly expose the hidden factors behind road accidents; but they cannot expose the trend of the road accidents in different locations. In this manuscript, we are demanding to elaborate road accident counts data to ascertain different districts where the trend of accident is increasing throughout the years. So that supplementary attention would be on these districts to overcome the accident trend. In order to do this, we have proposed a context to examine road accident time series data that uses both data mining and traditional statistical techniques. This charter inputs the road accident counts for different time series and then normalizes the time series. Further, it performs agglomerative clustering (AGNES) algorithm. Further a time series merging algorithm is proposed to find the representative time series (RTS) for each cluster. Finally, trend analysis is performed on every RTS of different clusters.

II. RELATED WORKS

Data mining has been proven as a trustworthy technique to scrutinize road accidents and deliver prolific results. Most of the road accident data scrutiny use data mining techniques, focusing on ascertaining factors that affect the severity of an accident. However, any mutilation resulting from road accidents is always undesirable in terms of health, property mutilation and other commercial factors. Sometimes, it is found that road accident occurrences are more numerous at certain specific locations. The analysis of these locations can help in perceiving certain road accident features that make a road accident to arise repeatedly in these locations. Association rule mining is one of the widespread data mining methods that ascertain the correlation in various attributes of road accident. In this paper, we first smeared k-means algorithm to group the accident locations into three categories, high-frequency, moderate-frequency and low-frequency accident locations.

K-means algorithm grosses accident frequency count as a parameter to cluster the locations. The association rule mining to characterize these locations. The rules revealed diverse factors associated with road accidents at diverse locations with fluctuating accident frequencies. The association rules for high-frequency accident location disclosed that intersections on highways are more hazardous for each form of accidents. High-frequency accident locations mostly involved two-wheeler accidents at hilly regions. In moderate-frequency accident locations, colonies adjacent to local roads and juncture on highway roads are originate hazardous for pedestrian hit accidents. Low-frequency accident locations are scattered throughout the district and the utmost of the accidents at these locations were not life-threatening. Although the data set was restricted to some selected attributes, our approach extracted some useful hidden evidence from the data which can be exploited to revenue some defensive efforts in these locations.

Identifying and profiling black spots and black zones in terms of accident related data and location characteristics must afford innovative perceptions into the complexity and causes of road accidents which, in fit, provide valuable input for government actions. In this paper, association rules are used to ascertain accident circumstances that recurrently arise composed at high frequency accident locations. Furthermore, these patterns are scrutinized and compared with recurrently arising accident characteristics at low frequency accident locations. The asset of this approach lies within the identification of relevant variables that make a strong contribution towards a better understanding of accident circumstances and the discriminating of eloquent accident patterns from more discriminating accident circumstances to profile black spots and black zones. The use of this data mining algorithm is predominantly beneficial in the context of large datasets on road accidents, since data mining can be defined as the extraction of information from large amounts of data.
Fallouts show that human and behavioral aspects are of great importance when scrutinizing recurring accident patterns. These factors play a significant role in identifying traffic safety problems in general. However, the most perceptive accident characteristics between high frequency accident locations and low frequency accident locations are mainly correlated to infrastructure and location characteristics. Unsupervised learning is an approach of learning where instances are automatically placed into eloquent groups based on their similarity. That paper familiarizes the vitalperceptions of unsupervised learning while it surveys the recent clustering algorithms. Moreover, recent advances in unsupervised learning, such as assemblages of clustering algorithms and distributed clustering, are described. Cluster analysis is an unsupervised learning method that organizes a cornerstone of an intelligent data analysis process. It is used for the exploration of inter-relationships surrounded by a collection of patterns, by unifying them into homogeneous clusters. It is called unsupervised learning because unlike classification (known as supervised learning), no a priori classification of some patterns is existing to use in labeling others and inferring the cluster structure of the entire data.

III. PROPOSED FRAMEWORK

A context is developed for exploration of road accident data. The different phases of framework are discussed below:

A. DATA PREPROCESSING

Data preprocessing is one of the imperativechore to be done prior to analyzing data. Data preprocessing leads to data cleaning such as removing noise, handling missing values and smearing various data transformations in order to get the data equipped for the analysis. Time series data also requires data preprocessing in order to get it normalized prior to the analysis. Normalization of time series prior to analysis assists in handling certain difficulties such as amplitude scaling, noise, offset translation etc. Therefore, preprocessing prior to the analysis helps in getting the precisefallouts. In order to normalize the time series data, z-score normalization method is used. Z-score method normalize a time series $T = \{t_1, t_2, t_3, ..., t_n\}$ into a normalized time series $NT = \{Nt_1, Nt_2, Nt_3, ..., Nt_n\}$ such that

$$\mu(NT) \approx 0 \text{ and } \sigma(NT) \approx 1$$

where $\mu(NT)$ and $\sigma(NT)$ are the mean and standard deviation respectively of normalized time series $NT$. The z-score formula for normalizing time series is given by Eq. 1.

$$NT = \sum_{i=1}^{n} \frac{(1 - \mu(T))}{\sigma(T)}$$

(1)

B. SIMILARITY MEASURE FOR TIME SERIES

A diversity of prevalent similarity measures are exist such as Euclidean distance, dynamic time warping (DTW), correlation coefficient and triangle distance metric (TDM) similarity measure for time series data. Similarity measure is very beneficial and vital component in clustering time series data. The outcome of similarity measure is a proximity square matrix of n dimension where n is the number of time series. In this study, we deliberate Euclidean distance, DTW, Pearson correlation coefficient (PCC) and TDM.

C. EUCLIDEAN DISTANCE

Euclidean distance is one of the prevalent and classic similarity measure used in several clustering algorithms such as K-means and hierarchical clustering. Euclidean distance can be demarcated as the distance concerning two points or vectors in Euclidean norm. Euclidean distance between two time series of identical length can be computed using Eq. 2 as follows:

$$\text{Dist}(T1, T2) = \sum_{i=1}^{n} \sqrt{(T1i - T2i)^2}$$

(2)

The above equation computes distance between two time series $T1$ and $T2$ of identical length of time sequence $n$.

D. DYNAMIC TIME WARPING

Dynamic time warping is alternative similarity measure for time series data which can measure distance concerning two time series objects even if their length is not similar. The advantage of DTW is that in order to minimize the distance concerning two time series $s_i = \{s_1, s_2, ..., s_n\}$ and $t_i = \{t_1, t_2, ..., t_n\}$ of length $n$ and $m$ respectively, it optimally align $s_i$ and $t_i$. The dynamic programming is used to discover the
similarity distance concerning two time series in matrix A. Matrix A is initialized to A[0,0] = 0 and A[i, j] = infinity, where i and j is not 0. The distance concerning two time series objects can be considered by recursively applying Eq. 3 as given below:

\[ A[i,j] = d(s_i, t_i) + \min \{ A[i,j-1], A[i-1,j], A[i-1,j-1] \} \]

where A is an n x m matrix in which Euclidean distance metric is used to discover the distance between s_i and t_i. The value of last cell A[n, m] symbolizes the distance concerning sequence s and t.

E. TRAINGLE DISTANCE METRIC

Triangle distance metric is alternative prevalent similarity measure for time series data which deliberates time series object as a vector in n dimensional space. Consider \( s_t = \{ s_{t1}, s_{t2}, \ldots, s_{tn} \} \) be an standardized time series object defined as follows:

\[ s_{tk} = \frac{\sum_{i=1}^{n} s_{ti}}{\sqrt{\sum_{i=1}^{n} s_{ti}^2}} \]  \hspace{1cm} (4)

The TDM concerning two vectors \( s_t \) and \( t_t \) can be premeditated using Eq. 4.

\[ d(s_t, t_t) = \frac{\sum_{k=1}^{n} (s_{tk} - t_{tk})^2}{\sqrt{\sum_{k=1}^{n} s_{tk}^2 \cdot t_{tk}^2}} = 1 - \sum_{k=1}^{n} s_{tk} \cdot t_{tk} \]

\[ \]  \hspace{1cm} (5)

TDM can be defined as the cosine of a triangle concerning two vectors and its value ranges concerning 0 and 2. The more different two time series objects are, the higher the value of TDM is. A low core symbolizes more similar time series objects while higher score symbolizes dissimilar time series objects.

F. HIERARCHICAL CLUSTER ANALYSIS

Cluster analysis primarily concerns with the assemblage of data objects into one or more groups. In clustering, the objects with similar properties are allotted to same group while data objects with different properties are allotted to different groups. There are numerous clustering algorithms such as partition based clustering, density based clustering etc. The choice of clustering algorithm rest on type and nature of data. There are various algorithms for clustering of time series data exist. Our approach recycled hierarchical clustering algorithm for clustering of time series data. Hierarchical clustering algorithm is of two types—agglomerative hierarchical clustering and divisive hierarchical clustering. The time complexity of first one is O (n^3) and the second one is O (2n), hence agglomerative clustering is sooner than divisive clustering. We used agglomerative hierarchical clustering algorithm (AGNES) in our study. The main obligation for a hierarchical clustering algorithm is the similarity measure which proceeds vital part in clustering process. In this study, we relate AGNES algorithm using seven versions consuming CPCC with Euclidean, TDM and DTW similarity measures. CPCC can be defined as a measure of the correlation concerning the Cophenetic distance of two time series objects and the original distance matrix.

G. TIME SERIES MERGING

Cluster analysis results in standardized segments of the time series data. Each cluster entails of several time series objects that are similar in nature. Hence there is a need to form a representative time series which can represent the entire time series. To discover the time series that can epitomize the entire cluster, a time series merging algorithm is formed, that takes DTW distance to compute the closest time series.

H. MULTI VIEW POINT BASED SIMILARITY MEASURE

In this fragment, rather than inspecting two documents similarity from the same cluster (the documents belongs to), inspecting in different viewpoints is deliberated. This will assists explore both inter cluster similarity and intra cluster similarity. Clustering is a useful method that categorizes a large quantity of unordered text documents into a small number of eloquent and articulate clusters, thereby providing a basis for natural and informative navigation and browsing mechanisms. There are some clustering methods which have to assume some cluster association between the data objects that they are applied on. Similarity between a pair of objects can be defined either explicitly or implicitly. The major difference concerning a traditional dissimilarity/similarity measure and ours is that the prior uses only single viewpoint, which is the origin, while the latter utilizes many diverse viewpoints,
which are objects assumed to not be in the same cluster with the two objects being restrained. Using multiple viewpoints, more informative assessment of similarity could be accomplished. Theoretical analysis and empirical study are steered to support this claim.

IV. TIME SERIES MERGING ALGORITHM

Time series is one of the prevalent data types that can be found in many domains such as business, medical, meteorological fields, etc. Ascertaining potential trends in time series is vital because it conveys knowledge about what has taken place in the past and what will take place in time to come. Trend analysis in the time series is the practice of collecting and endeavoring to spot patterns. Various data mining methods such as clustering, classification, regression, etc. can be used to expose those trends. In this work, we developed a context to scrutinize the time series data, which cluster time series affording to their similarity.

A. ALGORITHM: MERGING TIME SERIES

Input: n number of clusters
Output: A single representative time series for each cluster
Process:
Begin,
For i = 1 to n // i is cluster id
1. Calculate DTW distance between every time series in the cluster i
2. Merge the two closest time series objects (take the average of each data point)
3. Go to 1
4. Repeat until there is one time series object remains
5. Return a single time series to represent the cluster
End

V. EVALUATIONS AND DISCUSSION

As the road accident data consists of collection of road accident details happened at different locations at different time series. As these data’s has been taken to process the information. Initially, the data’s are separated based on location by using clustering technique. After clustering process, number of clusters is formed. With this help of clusters, the similarities between time series has found and the results reveals that road accident trend is going to increase in certain clusters and those districts should be the prime concern to take preventive measure to overcome the road accidents. In addition, for better clustering, data cleaning process such as stemming and stop word removal is applied. Then synonym word replacement is made which increases the relativity among the documents. Two dimensions are averaged into one dimension and so three dimension data is converted into two dimension data and then clustering is applied.

The following Table 1 describes the experimental result for existing system analysis. The table contains weight of text document.

<table>
<thead>
<tr>
<th>S.NO</th>
<th>Weight of Document</th>
<th>Weight of Clustering Document</th>
<th>Average of Clustering Document [%]</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>200</td>
<td>155</td>
<td>77.5</td>
</tr>
<tr>
<td>2</td>
<td>250</td>
<td>220</td>
<td>88.00</td>
</tr>
<tr>
<td>3</td>
<td>300</td>
<td>272</td>
<td>90.66</td>
</tr>
<tr>
<td>4</td>
<td>350</td>
<td>322</td>
<td>92.00</td>
</tr>
<tr>
<td>5</td>
<td>400</td>
<td>383</td>
<td>95.75</td>
</tr>
<tr>
<td>6</td>
<td>450</td>
<td>429</td>
<td>95.33</td>
</tr>
<tr>
<td>7</td>
<td>500</td>
<td>468</td>
<td>93.60</td>
</tr>
<tr>
<td>8</td>
<td>550</td>
<td>523</td>
<td>95.05</td>
</tr>
<tr>
<td>9</td>
<td>600</td>
<td>578</td>
<td>96.33</td>
</tr>
<tr>
<td>10</td>
<td>650</td>
<td>633</td>
<td>97.74</td>
</tr>
</tbody>
</table>

The experimental result for weight of clustering text document and average of text document clustering details are shown in the Fig 1 and Fig 2.

Fig 1 Existing - Average Clustering Documents
Road accidents are one of the chief issues for untimely death, partial or full disability and property damage, which is unacceptable in any form. A new document clustering method based on correlation preserving indexing is used which results in better correlation identification is achieved. Maximizes the correlation between the documents in the local patches and minimizes the correlation between the documents outside these patches. Considers both single view point and multiple viewpoints so that inter and intra cluster similarity can be analyzed effectively. As it is difficult and time consuming to scrutinize every time series of each cluster. A time series merging algorithm is also anticipated to merge all the time series and form a representative time series for each cluster. Finally, this representative time series algorithm is analyzed using least square regression method. In the future, it would also be possible to apply the clustering using hierarchical clustering algorithms so that intra and inter cluster similarity can be analyzed well. It would be interesting to explore how they work on other types of sparse and high-dimensional data.

VI. CONCLUSION

References