A Multivariate Time Series Clustering Approach for Crime Trends Prediction

B. Chandra
Indian Institute of Technology, Kanpur
Kanpur, India.
Permanent Address: IIT Delhi.
Email: bchandra104@yahoo.co.in.

Manish Gupta
Institute for Systems Studies and Analyses
Metcalfe House, Delhi, India.

M. P. Gupta
Indian Institute of Technology, Delhi
New Delhi, India.

***Abstract*—In recent past, there is an increased interest in time series clustering research, particularly for finding useful similar trends in multivariate time series in various applied areas such as environmental research, finance, and crime. Clustering multivariate time series has potential for analyzing large volume of crime data at different time points as law enforcement agencies are interested in finding crime trends of various police administration units such as states, districts and police stations so that future occurrences of similar incidents can be overcome. Most of the traditional time series clustering algorithms deals with only univariate time series data and for clustering high dimensional data, it has to be transformed into single dimension using a dimension reduction technique. The conventional time series clustering techniques do not provide desired results for crime data set, since crime data is high dimensional and consists of various crime types with different weightage. In this paper, a novel approach based on dynamic time wrapping and parametric Minkowski model has been proposed to find similar crime trends among various crime sequences of different crime locations and subsequently use this information for future crime trends prediction. Analysis on Indian crime records show that the proposed technique generally outperforms the existing techniques in clustering of such multivariate time series data.**

I. Introduction

In the present scenario, law enforcement agencies are facing difficulty in finding useful patterns from large volume of time series data of crime. Advanced analytical methods are required to extract useful information from large amount of crime data [9], [10]. Clustering techniques is looked upon as a solution to such problems. Clustering [1] as unsupervised technique is the process of organizing objects into groups such that similarity within the same cluster is maximized and similarities among different clusters are minimized. Unlike usual clustering methods, time series clustering dealt with the problem of sequential data patterns [6] and clusters the similar trends presents in the data.

Keogh et al. [7] has summarized time series clustering in two main categories i.e. whole clustering and subsequence clustering. Whole clustering is performed on many individual time series to group similar series into clusters whereas subsequence clustering is based on sliding window extractions of a single time series and aims to find similarity and differences among different time windows.

Traditional time series clustering methods [3], [8], [12], [16]–[19] can not be used for finding crime trends since crime data is high dimensional data and all crime types do not have equal weightage. In crime data, the objective is to cluster the crime location such as states, districts, and police stations with similar crime trends over multivariate time series data. In this paper, a novel approach for multivariate time series clustering technique based on dynamic time wrapping (DTW) [5] and parametric Minkowski model [4] has been proposed to find similar crime trends. Since all types of crime do not have equal weightage, for example, murder will have more weightage over kidnapping and hurt. DTW together with Parametric Minkowski model is used to consider the weightage scheme in the clustering algorithm. In parametric Minkowski model, the distance function is defined by a weighted version of the Minkowski distance measure. The parameters for this model are the weights in different dimensions. The effectiveness of the proposed approach has been illustrated in depth on Indian crime dataset provided by the Indian National Crime Records Bureau.

Section 2 of the paper highlights the major work done in the area of time series clustering and the limitation of those algorithms. The proposed approach along with dynamic time wrapping and parametric Minkowski model has been described in section 3. Section 4 shows the results of multivariate time series clustering of Indian crime data under various crime types such as murder, kidnapping etc. Conclusion is given in the last section of the paper.

II. Related Work

In this section, some of the widely known time series clustering algorithms have been described in brief with their limitation.

Clustering of timer series data has two major limitations such as handling of high dimensional data and handling missing value. Wang et. al. [18] has proposed characteristic-based clustering for time series data based on their structural characteristics such as trend, seasonality, periodicity, serial correlation, skewness, kurtosis, chaos, nonlinearity, and self-similarity. In this approach, measurements of characteristics take so much effort and make time series clustering process more time consuming. For crime data, measurement of such characteristics is not possible because of paucity of data available.

1-4244-2384-2/08/$20.00 © 2008 IEEE

Wang et. al. [14] proposed a dimensionality reduction technique i.e. piecewise vector quantized approximation [2] for time series analysis that significantly improves the efficiency and accuracy of similarity searches. Such a technique is mostly suitable on text summarization but not in the crime domain.

Singhal et. al. [3] has proposed a methodology based on calculating the degree of similarity between multivariate time-series datasets using two similarity factors. One similarity factor is based on principal component analysis and the angles between the principal component subspaces while the other is based on the Mahalanobis distance between the datasets. The methodology has the basic limitations of PCA and fails for the data with no correlation among various dimensions.

Xiong et. al. [19] has proposed a model-based approach using mixtures of autoregressive moving average (ARMA) models [13]. An expectation-maximization (EM) algorithm has been used for learning the mixing coefficients as well as the parameters of the component models and also applied Bayesian information criterion (BIC) to determine the number of clusters in the data. The model requires large number of sequential data but in crime domain this data is hardly available.

To overcome the limitations of the existing algorithm, a novel approach based on dynamic time wrapping with parametric Minkowski model has been proposed to find similar crime trends and subsequently use this information for future crime trends prediction. The proposed approach is described in the next section.

III. Proposed Approach

This section gives the proposed approach of crime trends prediction in details.

A. Methodology

Euclidean distance measure is commonly used for non-time series data clustering but it is not suitable for multivariate time series clustering. Instead of Euclidean distance measure in stand alone mode, dynamic time wrapping (DTW) [5] provides better results for many application areas of time series. The major limitation of DTW is of handling sequential multivariate data with different weights of each dimension. So, the existing DTW is also not sufficient for tackling the problem of prediction of crime trends since crime sequences are having property of multidimensional with varying weights. To overcome this limitation of DTW, parametric Minkowski model [4] is introduced in measuring the distance matrix required for DTW. Firstly, DTW using parametric Minkowski model is measured for each pair of districts and then hierarchical clustering with single linkage method i.e. nearest neighbor method has been applied to find the similar crime trends and subsequently crime trends prediction. Hierarchical clustering algorithm [15] is preferred to partitioning clustering methods like k-means [11] for time series data because time series data can be visualized using dendrogram. The partition methods has also not been considered suitable since numbers of clusters are unknown for crime data.

Hence, the proposed approach suggests the necessary modification in DTW and introduce weightage scheme as it is mandatory for crime trends prediction. To demonstrate the utility of proposed approach, a brief overviews of dynamic time wrapping (DTW) and parametric Minkowski model are given in subsequent subsections.

B. Dynamic Time Wrapping (DTW)

Given two time series, $X = x_1, x_2, \ldots, x_i, \ldots, x_n$ and $Y = y_1, y_2, \ldots, y_j, \ldots, y_m$, DTW aligns the two series so that their difference is minimized. To this end, a distance matrix D of order $n \times m$, where the (i, j) element of the matrix D contains the distance $d(x_i, y_j)$ between two points x_i , and y_j. The Euclidean distance is normally used. A wrapping path, $W = w_1, w_2, \ldots, w_k, \ldots, w_K$, where $max(m, n) \leq K \leq m + n - 1$ that has the minimum distance between the two series is of interest It is a set of matrix elements that satisfies three constraints:

1) Boundary condition i.e. , w_1 = first element of distance matrix D and w_K = last element of distance matrix D.
2) Continuity restricts the allowable steps to adjacent cells, and
3) Monotonicity forces the points in the warping path to be monotonically spaced in time.

Mathematically,

$$DTW = min\frac{\sum_{k=1}^{K} w_k}{K} \tag{1}$$

Dynamic programming is generally being used to effectively find this path by evaluating the following recurrence, which defines the cumulative distance as the sum of the distance of the current element and the minimum of the cumulative distances of the adjacent elements.

$$d_c(i, j) = d(x_i, y_j) + min\{d_c(i-1, j-1), d_c(i-1, j), d_c(i, j-1)\} \tag{2}$$

The major advantage of the DTW is that two sequences need not to be of same length.

C. Parametric Minkowski Model

In parametric Minkowski model [4], the distance function is defined by a weighted version of the Minkowski distance measure. The parameters for this model are the weights in different dimensions. Let the weights for the d dimensions are denoted by $\lambda_1, \ldots, \lambda_d$. For a pair of data objects $X = x_1, \ldots, x_d$ and $Y = y_1, \ldots, y_d$, the parametric distance is defined as follows

$$f(X, Y, \lambda_1, \ldots, \lambda_d) = (\sum_{i=1}^{d} \lambda_i \|(x_i - y_i)\|^p)^{1/p} \tag{3}$$

A higher value of λ_i indicates a greater significance for dimension i. In this model, the values of these weights are computed using p-norm mean square error function. The error function is also contained user defined similarity values for the different data points. The gradient search method is further applied to solve the subsequent mathematical equations. Therefore, it is difficult to determine the most suitable weights in the parametric Minkowski model as similarity values for the different data points are hard to be determined. For the current analysis, domain specific weight matrix is defined by the team members of Indian National Crime Records Bureau (NCRB).

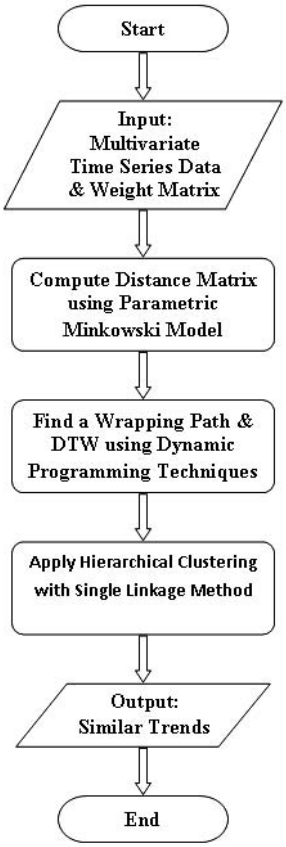

Fig. 1. Flowchart of Proposed Approach

D. Procedure

From the earlier discussion, the step be step procedure of proposed methodology can be summarized as follows,

1) Consider multivariate time series data as input
2) Choose weight matrix i.e. weights for different dimensions
3) Compute distance matrix D for each pair of time series using parametric Minkowski model
4) Find a wrapping path using dynamic programming techniques
5) Find DTW for each pair of time series
6) Apply hierarchical clustering with single linkage method for finding similar crime trends
7) Predict crime trends based on the similar crime trends obtained in step (6).

The flowchart of the methodology is given in Fig. 1. The effectiveness of the proposed approach has been illustrated on Indian crime records. The results and discussion is shown in the next section.

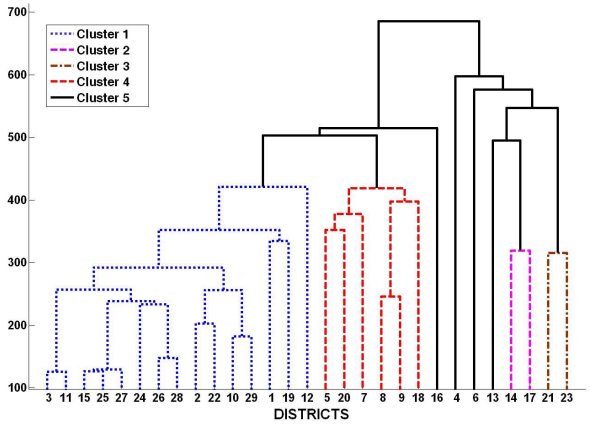

Fig. 2. Dendrogram of Crime Against Body for 2002-2006 using DTW with Euclidean

IV. Results and Discussion

In this paper, the proposed approach has been applied on the Indian crime dataset provided by National Crime Records Bureau (NCRB). It is important to know similar crime trends so that crime trends can be predicted in multivariate time series of crime. It is needed to describe briefly the Indian police system to understand nature of crime data. Indian constitution assigns responsibility for maintaining law and order to the states and union territories (UT), and almost all routine policing, including apprehension of criminals, is carried out by state-level police forces. India is divided into 28 states and 7 union territories (UT). Prediction of crime trends of various police administration units has remained a constant area of governmental concern since these states and UT are having diversities in area, population and crime density. So, The proposed approach effectively applied to identify similar crime trends with similar crime density and predict the crime trends for future implications.

Indian crime data is provided by National Crime Records Bureau (NCRB) for prediction of crime trends. Crime data contains the crime records of 29 districts of an Indian state for 2002-2006 under seven important crimes i.e. murder, attempt to murder, abduction, kidnapping, assault, hurt and culpable homicide not amounting to murder. Since, all types of crime do not have equal weights, for example, murder will have more weights over kidnapping and hurt. To incorporate weightage schemes for different crime types, parametric Minkowski model has been used to calculate dynamic time wrapping(DTW) amongst 29 districts of the state. Hierarchical clustering with single linkage method i.e. nearest neighbor method has been further applied for finding similar crime trends. The threshold value has been fixed 70% of the maximum DTW. Fig. 2 and Fig. 3 are given to compare the results of the existing approach i.e. DTW with Euclidean and the proposed approach i.e. DTW with parametric Minkowski model.

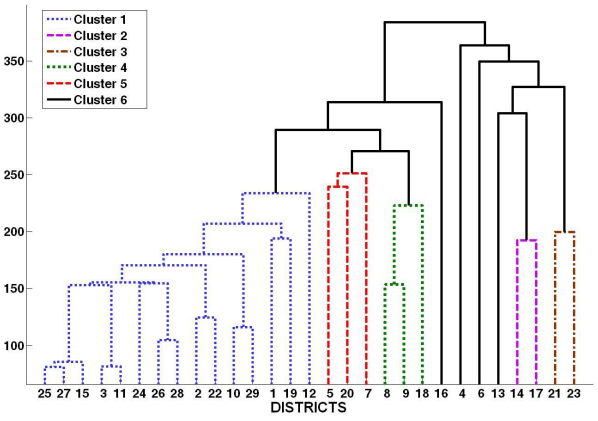

Fig. 3. Dendrogram of Crime Against Body for 2002-2006 using DTW with Parametric Minkowski Model

Fig. 2 shows the results of crime trends clustering using Dynamic Time Wrapping (DTW) Euclidean for 2002-2006. Different colors show different cluster groups with similar crime trends. From the dendrogram given in Fig. 2, five clusters with similar crime trends have been formed. First cluster contains 15 districts i.e. district number 3, 11, 15, 25, 27, 24, 26, 28, 2, 22, 10, 29, 1, 19 and 12 with similar crime trends. Districts 5, 20, 7, 8, 9 and 18 formed the second cluster group with similar crime trends. District 14 and 17 are also having similar crime trends whereas district 21 has similar crime trends with district 23. The remaining four districts i.e. 16, 4, 6 and 13 did not match with the crime trends of other districts.

Fig. 3 shows the results of crime trends clustering using Dynamic Time Wrapping (DTW) with parametric Minkowski model for 2002-2006. For this application, NCRB experts provide the requisite weight matrix as [1, .9, .7, .6, .5, .3, .6, .7], which means murder should be given maximum weights as compared to other crime types. The crime type ”hurt” is given only 30% weightage as compared to murder. DTW with Euclidean distance considers equal weightages for both the crime viz. ”murder ” and ”hurt”, which is not a desirable for crime time series clustering. By introducing weightage scheme, second cluster has been split into two different clusters.

In DTW with parametric Minkowski model, the order of similarity has also been changed in the first cluster. Instead of the similarity order of districts 3, 11, 15, 25, 27 in Fig. 2, the order of similar crime trends in Fig. 3 has become as districts 25, 27, 15, 3, 11. The order of similarity is significant since it decides for a district, which are the most similar districts in terms of crime.

Fig. 4 and Fig. 5 show time series plot of ”kidnapping” type of crime for two districts i.e. District-25 and District-27 with similar crime trends. Similarly, Fig. 6 and Fig. 7 show similar crime trends for District-25 and District-27 for ”hurt” type of crime. It can be observed from Fig. 4 and Fig. 5 that District-

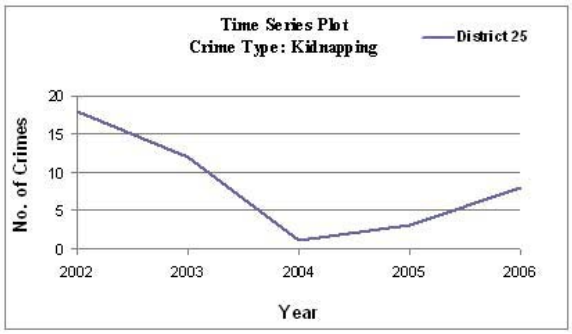

Fig. 4. Time Series Plot of District 25: Crime Type-Kidnapping

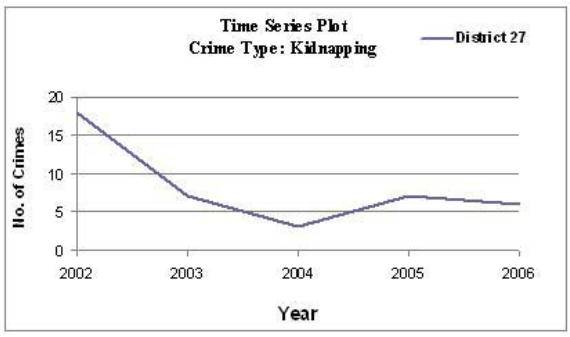

Fig. 5. Time Series Plot of District 27: Crime Type-Kidnapping

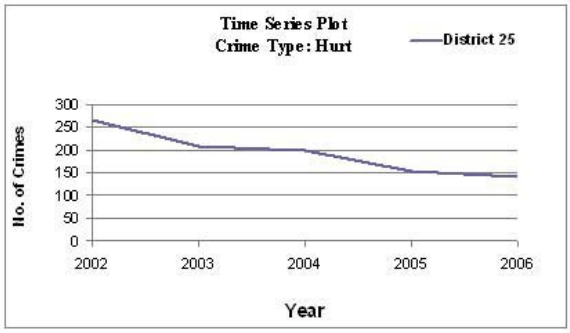

Fig. 6. Time Series Plot of District 25: Crime Type-Hurt

25, and District-27 have similar crime trends for kidnapping crime. Similarly, Fig. 6 and Fig. 7 show the similar crime trends for District-25 and District-27 for hurt crime. To observe the change in crime trends, the crime data of period 1995-2006 has been split into three block periods i.e. 1995-1998, 1999-2002 and 2003-2006. The following interesting results have been noticed by analysis of results of these three block periods.

1) Crime trends in the districts 3 and 11 are most similar for all block periods 1995 -1998, 1999-2002, 2003-2006.
2) The crime trends of the districts 1, 2, 22, 26, 19, 10, 5, 7,

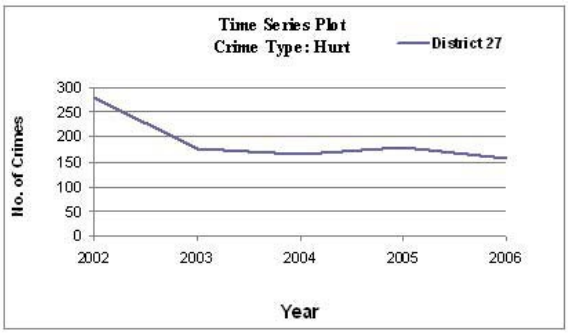

Fig. 7. Time Series Plot of District 27: Crime Type-Hurt

8, 9, 10, 29 and 20 are grouped into same cluster for all the block periods 1995 -1998, 1999-2002, 2003-2006.

From the above results, it is found that the proposed methodology exhibits the desired results for prediction of crime trends. The proposed approach is useful and can easily be applied in those types of time series clustering problems where each dimension has different significance.

V. Conclusion

In this paper, a novel approach for multivariate time series clustering based on dynamic time wrapping and parametric Minkowski model has been proposed for finding similar crime trends efficiently and subsequently predict crime trends. The effectiveness of the proposed approach over the existing clustering algorithms has been illustrated using Indian crime data. The approach can play an important role for wider variety of multivariate time series clustering problems especially where the dimensions do not have equal weightages.

Acknowledgements

We are highly indebted to Shri Sudhir Awasthi, Director, National Crime Records Bureau (NCRB) for funding a project on crime data mining. We are also thankful to him and his team members for sharing their valuable knowledge in generating weights for carrying out clustering of crime data using proposed approach.

References

[1] A. K. Jain, M. N. Murty, P. J. Flynn, *Data clustering: a review*, ACM Computing Surveys, 31(3), 1999, 264-323.

[2] A. Lendasse, D. Francois, V. Wertz, M. Verleysen, *Vector quantization: a weighted version for time-series forecasting*, Future Generation Computer System 21 (7), 2005, 1056-1067.

[3] A. Singhal, D. E. Seborg, *Clustering multivariate time-series data*, Journal of Chemometrics, 19, 2005, 427-438.

[4] C. Aggarwal, *Towards systematic design of distance functions for data mining applications*, ACM KDD Conference, 2003.

[5] D. Berndt, J. Clifford, *Using dynamic time warping to find patterns in time series*, AAAI-94, Workshop on Knowledge Discovery in Databases, 1994, 229-248.

[6] E. Keogh, *The UCR Time Series Data Mining Archive*, http://www.cs.ucr.edu/~eamonn/TSDMA/index.html. Computer Science & Engineering Department, University of California, Riverside, CA, 2002.

[7] E. Keogh, J. Lin, W. Truppel, *Clustering of time series subsequences is meaningless: Implications for past and future research*, In Proceedings of the 3rd IEEE International Conference on Data Mining, Melbourne, FL, USA, 2003, 115-122.

[8] E. Keogh, S. Kasetty, *On the Need for Time Series Data Mining Benchmarks: A Survey and Empirical Demonstration*, In proceedings of the 8th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. Edmonton, Alberta, Canada, July 23-26, 2002, 102-111.

[9] H. Chen, D. Zeng, H.Atabakhsh, W. Wyzga, J. Schroeder, *COPLINK: managing law enforcement data and knowledge*, Communications of the ACM 46 (1), 2003, 28-34.

[10] J. J. Corcoran, I. D.Wilson, J. A.Ware, *Predicting the geo-temporal variations of crime and disorder*, International Journal of Forecasting 19, 2003, 623-634.

[11] J. McQueen, *Some methods for classification and analysis of multivariate observations*, Proceedings of Symposium on Mathematics, Statistics And Probability, 5^{th}, Berkeley, 1, 1967, 281-298.

[12] L. Hansheng, V. Govindaraju, *Generalized regression model for sequence matching and clustering*, Knowledge and Information System, 12(1), 2007, 77-94.

[13] M. Corduas, D. Piccolo, *Time series clustering and classification by the autoregressive metric*, Computational Statistics & Data Analysis 52, 2008, 1860 - 1872.

[14] Q. Wang, V. Megalooikonomou, *A dimensionality reduction technique for efficient time series similarity analysis*, Information Systems 33, 2008, 115-132.

[15] S. C. Johnson, *Hierarchical clustering schemes*, Psychometrika, 32(3), 1967, 241-254.

[16] S. Guha, N. Mishra, R. Motwani, L. O'Callaghan, *Clustering Data Streams*, In proceedings of the 41^{st} Annual Symposium on Foundations of Computer Science. Redondo Beach, CA, Nov 12-14. 2000, 359-366.

[17] T. W. Liao, *Clustering of time series data:a survey*, Pattern Recognition, 38, 2005, 1857-1874.

[18] X. Wang, K. Smith, *Characteristic-Based Clustering For Time Series Data*, Data Mining And Knowledge Discovery, 13, 2006, 335-364.

[19] Y. Xiong, D. Y. Yeung, *Time series clustering with ARMA mixtures*, Pattern Recognition, 37, 2004, 1675 - 1689.