International Journal of Advanced Engineering Research and Science (IJAERS) *[Vol-3, Issue-11, Nov- 2016]*
https://dx.doi.org/10.22161/ijaers/3.11.27 *ISSN: 2349-6495(P) | 2456-1908(O)*

A Survey On Outlier Detection Methods In Spatio-Temporal Datasets

M L Prasanthi[1], A Krishna Chaitanya[2], Dr.N Sambasiva Rao[3]

[1]Assistant Professor in Information Technology, BVRITH College of Engineering for Women, Hyderabad, JNTUH, India.
[2]Assistant Professor in Information Technology,Institute of Aeronautical Engineering, Hyderabad, JNTUH, India.
[3]Professor in Computer Science and Engineering, SRIT for Women, Warangal, JNTUH, India.

Abstract*—Outlier mining has many applications in the real world, such as Weather forecasting, Traffic management, Forest fire, and crop sciences. Extending these applications to Satellites, sensor networks, RFID technology, GPS and telecommunication systems which have become centers for gathering of large sources of data, several interesting facts from Spatio temporal datasets can be extracted. This paper summarizes recent works on the different outlier detection methods which are suitable to detect outliers from Spatio temporal datasets.*

Keywords—spatio temporal data, outlier detection.

I. INTRODUCTION

Outlier detection is a emerging field, which has lot of significance due to the increasing amount of spatio-temporal data available, and the need to understand and interpret it. A spatiotemporal object can be defined as an object that has at least one spatial and one temporal property. The spatial properties are location and geometry of the object. The temporal property is timestamp or time interval for which the object is valid. The spatio-temporal object usually contains spatial, temporal and thematic or non-spatial attributes. Examples of such objects are moving car, forest fire, and earth quake. Spatiotemporal data sets essentially capture changing values of spatial and thematic attributes over a period of time. An event in a spatio temporal dataset describes a spatial and temporal phenomenon that may happens at a certain time t and location x. Examples of event types are earth quake, hurricanes, road traffic jam and road accidents. In real world many of these events interact with each other and exhibit spatial and temporal patterns which may help to understand the physical phenomenon behind them. Therefore, it is very important to identify efficiently the spatial and temporal features of these events and their relationships from large spatio temporal datasets of a given application domain.

Spatio-temporal data mining is the discovery of interesting spatial patterns from data over time using data mining techniques on spatially and temporally distributed data. One such pattern is a spatio-temporal outlier. A spatio-temporal outlier is a spatio-temporal object whose thematic (non-spatial and non-temporal) attributes are significantly different from those of other objects in its spatial and temporal neighborhoods.

This paper is organized as follows section 2 emphasis on classification of outlier detection. Section 3 focuses on taxonomy of approaches to collect Spatio-temporal data. Section 4 summarizes various algorithms to detect outliers from multivariate spatio temporal data

II. CLASSIFICATION OF OUTLIER DETECTION METHODS

The most significant approaches for outlier detection are included in Figure1:

1) Distribution-based approaches that make use of standard statistical distribution to model the data declaring as outliers the objectsthat deviate from the model.

2) Depth-based techniques which are based on computational geometry and compute different layers of convex hulls declaring as outliers the objects belonging to the outer layers

3) Distance-based approaches which compute the proportion of database objects that are at specified distance from a target object

4) Density-based approaches which assign a weight to each sample based on their local neighborhood density.

https://dx.doi.org/10.22161/ijaers/3.11.27 *ISSN: 2349-6495(P) | 2456-1908(O)*

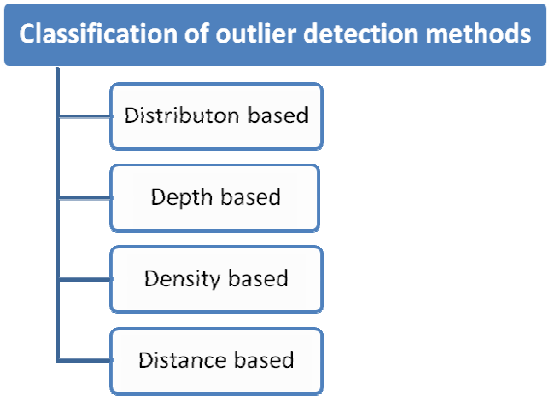

Fig.1: Classification of outlier detection methods

A different classification is based on the outlier detection output and divides into: Labeling and Scoring techniques. Labeling methods partition the data into two non-overlapping sets (outliers and non-outliers) and scoring methods offer a ranking list by assigning to each datum a factor reflecting its degree of outlierness. These former methods exploit a hard decision about the sets, the latter ones deal with a sort of soft decision about the membership of each datum to the set.

Another way of classifying Outlier detection methods can be divided between univariate methods, proposed in earlier works in this field, and multivariate methods that usually form most of the current body of research.

Another fundamental taxonomy of outlier detection methods is between parametric (statistical) methods and nonparametric methods that are model-free (e.g., see (Williams et al., 2002)).

Statistical parametric methods either assume a known underlying distribution of the observations or, at least, they are based on statistical estimates of unknown distribution parameters. These methods considers outliers as those observations that deviate from the model assumptions. They are often unsuitable for high-dimensional data sets and for arbitrary data sets without prior knowledge of the underlying data distribution. Within the class of non-parametric outlier detection methods one can set apart the data-mining methods, also called distance-based methods. These methods are usually based on local distance measures and are capable of handling large databases .

Another related class of methods consists of detection techniques for spatial outliers. These methods search for extreme observations or local instabilities with respect to neighboring values, although these observations may not be significantly different from the entire population.

Univariate Statistical Methods: Most of the earliest univariate methods for outlier detection rely on the assumption of an underlying known distribution of the data, which is assumed to be identically and independently distributed .Moreover, many discordance tests for detecting univariate outliers further assume that the distribution parameters and the type of expected outliers are also known (Barnett and Lewis, 1994). Needless to say, in real world data-mining applications these assumptions are often violated. A central assumption in statistical-based methods for outlier detection, is a generating model that allows a small number of observations to be randomly sampled from distributions G1,. . . , Gk, differing from the target distribution F, which is often taken to be a normal distribution N ¡ µ, σ2 ¢ (see (Ferguson, 1961; David, 1979; Barnett and Lewis, 1994; Gather, 1989; Davies and Gather, 1993)). The outlier identification problem is then translated to the problem of identifying those observations that lie in a so-called outlier region.

III. TAXONOMY OF APPROACHES TO COLLECT SPATIO-TEMPORAL DATA

Two approaches are used to collect spatio temporal data.

1. Lagrangian approach
2. Eularian approach.

1. Lagrangian approach:

The Lagrangian approach is individual-based and entails tracking a specific individual. Figure 2 below, shows classification of various technologies used in lagrangian approach. The technologies used for the Lagrangian modeling are, on the other hand invasive, in form of a mark or device which is fitted on the object. They are designed specifically to retrieve specific spatio-temporal data type with high quality.

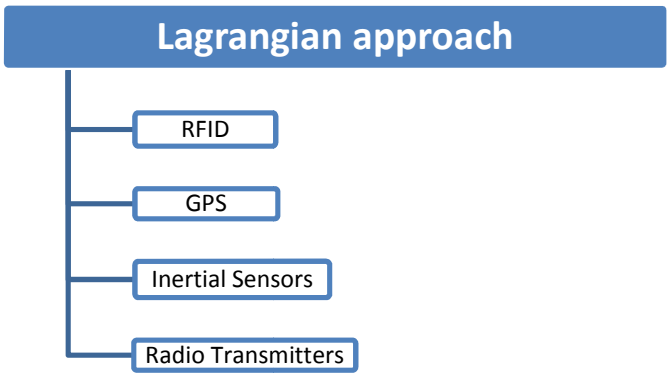

Fig.2: Lagrangian approach

https://dx.doi.org/10.22161/ijaers/3.11.27 *ISSN: 2349-6495(P) | 2456-1908(O)*

Eulerian approach:

The Eulerian approach is place-based and deals with the probability of presence of an individual or a group in a place and the change of this occurrence over time. Figure3 shows classification of various technologes used in Eularian approach. Various types of data can be retrieved from sensor technologies used for the Eulerian modeling, these data are subject to error and noise.

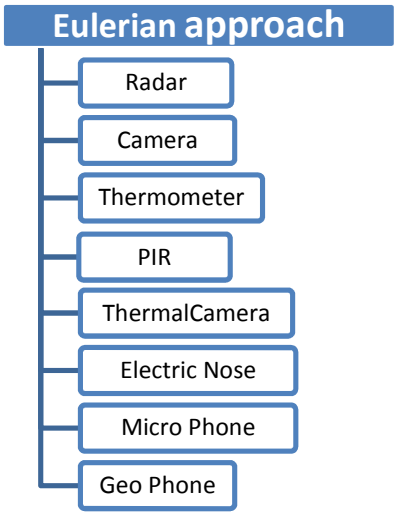

Fig.3: Eulerian approach

IV. OUTLIER DETECTION METHODS SUITABLE TO MULTIVARIATE SPATIO TEMPORAL DATA

Spatiotemporal data mining is used in various applications domains like Meterology,crop sciences,Forestry,Medicine,Geo physics, ecology, Transportation.Outlier detection in spatio temporal data patterns used to identify interesting abnormal patterns in huge datasets . some of such interesting patterns are detecting tropicalcyclonic paths, patterns of growing cancer cells in the cancer data of a patient in different time intervals.

In anomaly detection we also have the additional problem of identifying what is an anomaly in the domain: this is especially hard when dealing with multivariate data, because characterizing anomalies involving multiple features may not be intuitive and in most cases visualization approaches can't be used with high dimensionality. Finally, in multivariate data some features may not be available as frequently as others due to coming from different sources, leading to missing data concentrated in some features. The strategies for dealing with missing data in multivariate sets need to account for correlation between features.

Dataset composed of many correlated attributes or features are common in real world application. Traditional time series anomaly detection algorithms such as CUSUM or STCOD analyze only one feature of the dataset.While many anomalies may be detected by analyzing a single feature, other anomalies might affect multiple features without being evident in a single one. In this situation the analysis must be performed on multiple features at once. Univariate analysis also does not account for the possibility of features being correlated to each other; this information could be used, for example, to predict the behavior of a feature by analyzing other correlated features. Secondly, multivariate data is often heterogeneous; when types of features are different the analysis needs to rely on techniques that are general enough to be applicable to every data type involved. Other than the data type, different features may be differently distributed, or have different semantic (for example only positive are meaningful, only integer etc.).The survey of outlier detection algorithm shows that no algorithm can guarantee all our desired properties when applied to multivariate spatiotemporal data.

This paper summarizes few outlier detection approches used to identify outlier patterns from different multivariate spatio temporal datasets of various domains.

a) ST-SNN Algorithm:

K.P. Agrawal, Sanjay Garg and Pinkal Patel suggested Spatio-Temporal Shared Nearest Neighbor (ST-SNN) clustering approach, SNN which is based on the Shared Nearest Neighbor Similarity and modified version of existing density based clustering approach. It focuses on the clustering technique to detect outliers, which works well for high dimensional, arbitrary shaped, size and different density dataset and it is capable to handle high dimensional spatio-temporal data having different sizes and densities and also identifies arbitrary shaped clusters.

They experimented the algorithm on spatio-temporal dataset containing NDVI (Normalized difference vegetation index)values for different states in India. These NDVI values shows low vegetation when values are between 0.1 and 0.3 and shows high vegetation when values are between 0.8 and 1. Dataset has the columns: Grid code for each grid, Latitude of the grid, Longitude of the grid, state, City and 23 columns for NDVI values, where each column contains 16 days composite NDVI values.

In particular, algorithm first finds k-nearest neighbors for each data objects and then, as in SNN clustering approach, shared nearest neighbors are found between the pairs of points in terms of how many nearest neighbors the two points share. Using ST-SNN clustering, core points are

https://dx.doi.org/10.22161/ijaers/3.11.27 *ISSN: 2349-6495(P) | 2456-1908(O)*

identified in spatiotemporal data and clusters are created around the core points. The use of Shared Nearest Neighbor approach solves the problems with variable density and the unreliability of distance measure in high dimensions, while the use of core points handles problems with shape and size. The average runtime complexity of ST-SNN outlier detection algorithm is O (n2), where n is the number of objects in the dataset. The space complexity of the algorithm is also O (n2), since we need to store k-nearest neighbors which takes O (k*n) space and shared nearest neighbors similarity matrix which takes O(n2)

b) MUSTF Algorithm:

Gianluca Goffredi proposed a new algorithm "Multivariate Spatio-Temporal Anomaly Detection using Fisher's method(MuSTF) " to detect outliers in mobile network.MuSTF algorithm proved in detecting anomalies in the same data. Another particularly important improvement of MuSTF is the shorter delay in detection: our algorithm is able to detect the same anomaly several hours sooner than the state-of-the-art algorithms. Finally, MuSTF performance is less sensitive to small changes in the user-set parameters than other state-of-the-art algorithm, which relieves the user from having to discover the best settings to find acceptable results. We also showed how MuSTF is able to identify the spatial cluster that is involved in the anomaly, distinguishing clusters of nodes that belong to areas with different behavior even when close to each other.

c) COVSRE Method:

Alka Bhushan , Monir H. Sharker, Hassan A. Karimi suggested a statistical outlier detection method COVSRE(covariance free squared reconstruction error) which is suitable for spatio temporal datastreams. It uses the concept of incremental Principal component Analysis. The dataset used in their method is the air quality index (AQI) dataset which is publicly available from central Environmental Protection Agency (EPA) repository in USA (EPA, 2011).

EPA has placed sensors to measure pollutants across locations all over USA. The data is collected on hourly basis. Each sensor measures air pollutants at regular intervals and sends the measurement to the central data repository. AQI measures the quality of air which is computed based on the quantity of pollutants measured at each location at each given instance. From amongst 3000 sensors, 81 sensors from one geographically chosen area are selected for the experiments. The computational time to detect outliers using COVSRE method is O(nk).

For spatial applications, however, these global forms can only detect outliers in a non-spatial manner. This can result in false positive detections, such as when an observation's spatial neighbors are similar, or false negative detections such as when its spatial neighbors are dissimilar. To avoid mis-classifications, we demonstrate that a local adaptation of various global methods can be used to detect multivariate spatial outliers. In particular, we account for local spatial effects via the use of geographically weighted data with either Mahalanobis distances or principal components analysis. Detection performance is assessed using simulated data as well as freshwater chemistry data collected over all of Great Britain. Results clearly show value in both geographically weighted methods to outlier detection.

Derya Birant, Alp Kut proposed a new approach to find Spatio temporal outliers in large databases, the approach uses a 3-step algorithm constitutes clustering, checking spatial neighbors to identify spatial outliers, and checking temporal neighbors to identify spatio-temporal outliers.

d) ROSE(Rough Outlier set Extraction) Algorithm:

Alessia Albanese, Sankar K. Pal, and Alfredo Petrosino proposed a new rough set based approach called ROSE (Rough outlier set exraction) to detect spatiotemporal outliers. This method uses a new set called Kernel set as input instead of the universe, which is derived from universe U and detects outliers with low computational time.

e) Multi scale approach to detect spatio temporal outliers:

Tao Cheng ,Zhilin LiA used a multiscale approach to detect spatio temporal outliers ,where the detection process constitutes four steps classification, aggregation, comparision, and verification. The dataset used is Ameland, a barrier island in the north of the Netherlands, was chosen as a case study area. The process of coast change involves the erosion and accumulation of sediments along the coast, which is scale-dependent in space and time. It can be monitored through the observation of annual changes of landscape units such as foreshore, beach and foredune. The data set we used covers part of the island. The DEMs of six consecutive years (from 1989-1995) is displayed in Figure 2. It is hard to identify the outliers in the images displayed in Figure 2. The purpose of our experiment is to use the multiscale approach to detect the outliers in these six year DEMs.

V. CONCLUSION

The Spatio temporal outlier detection is very much essential to find out the noise in dataset. These outliers must be checked whether they are true or false outliers since the elimination of false outliers may affect the final analysis

https://dx.doi.org/10.22161/ijaers/3.11.27 *ISSN: 2349-6495(P) | 2456-1908(O)*

results and the presence of true outliers also makes confusion. In any datamining process the elimination of inconsistent values or outliers itself makes the process easier for further analysis and in spatial datamining it is very much essential since abundant data is involved in the processing. This paper discusses the approaches, methods and some algorithms used for outlier detection in spatio temporal data streams.

REFERENCES

[1] K.P. Agrawal, Sanjay Garg andPinkal Patel "Spatio-Temporal Outlier Detection Technique" IJCSC Vol 6 Number 2 April - Sep 2015

[2] Raymond T. Ng and Jiawei Han, CLARANS: A Method for Clustering Objects for Spatial Data Mining 2002.
Alka Bhushan , Monir H. Sharker, Hassan A. Karimi "INCREMENTAL PRINCIPAL COMPONENT ANALYSIS BASED OUTLIER DETECTION METHODS FOR SPATIOTEMPORAL DATA STREAMS"
Remote Sensing and Spatial Information Sciences, Volume II-4/W2, 2015 International Workshop on Spatiotemporal Computing, 13–15 July 2015.

[3] Derya Birant, Alp Kut "Spatio-Temporal Outlier Detection in Large Databases" Journal of Computing and Information Technology - CIT 14, 2006.

[4] Tao Cheng Zhilin Li "A MULTISCALE APPROACH TO DETECT SPATIAL-TEMPORAL OUTLIERS".

[5] Paul Harris Chris Brunsdon, Martin Charlton, Steve Juggins, Annemarie Clarke "Multivariate Spatial Outlier Detection Using Robust Geographically Weighted Methods".

[6] Gianluca Goffredi "MULTIVARIATE SPATIO-TEMPORAL ANOMALY DETECTION IN A MOBILE NETWORK" 2014-2015.

[7] V. Barnett and T. Lewis 1994 "Outliers in Statistical Data," John Wiley & Sons.

[8] P. Burge and J. Shaw-Taylor 1997 "Detecting Cellular Fraud Using Adaptive Prototypes," Proc. AI Approaches to Fraud Detection and Risk Management, pp. 9-13.

[9] T. Cover and J.A. Thomas 1991 "Elements of Information Theory," Wiley-International.

[10] T. Fawcett and F. Provost 1999 "Activity Monitoring: Noticing Interesting Changes in Behavior," Proc. ACM-SIGKDD Int'l Conf. Knowledge Discovery and Data Mining, pp. 53-62.

[11] S.B. Guthery, "Partition Regression 1974," J. Am. Statistical Assoc., vol. 69, no. 348, pp. 945-947.

[12] D.M. Hawkins 1976 "Point Estimation of Parameters of Piecewise Regression Models," J Royal Statistical Soc. Series C, vol. 25, no. 1, pp. 51-57.

[13] M. Huskova, 1993 "Nonparametric Procedures for Detecting a Change in Simple Linear Regression Models," Applied Change Point Problems in Statistics.

[14] G. Kitagawa and W. Gersch 1996 "Smoothness Priors Analysis of Time Series," Lecture Notes in Statistics, vol. 116, Springer-Verlag.

[15] E.M. Knorr and R.T. Ng 1998 "Algorithms for Mining Distance-Based Outliers in Large Data Sets," Proc. 24th Very Large Data Bases Conf.,pp. 392-403.

[16] U. Murad and G. Pinkas 1999 "Unsupervised Profiling for Identifying Superimposed Fraud," Proc. Third European Conf. Principles and Practice of Knowledge Discovery in Databases, pp. 251-261.

[17] R.M. Neal and G.E. Hinton 1993 "A View of the EM Algorithm that Justifies Incremental, Sparse, and Other Variants".

[18] T. Ozaki and G. Kitagawa 1995 "A Method for Time Series Analysis," Asakura Shoten, (in Japanese).

[19] J. Rissanen 1996 "Fisher Information and Stochastic Complexity," IEEE Trans. Information Theory, vol. 42, no. 1, pp. 40-47.

[20] K. Yamanishi, J. Takeuchi, G. Williams, and P. Milne May 2004 "Online Unsupervised Outlier Detection Using Finite Mixtures with Discounting Learning Algorithms," Data Mining and Knowledge Discovery J., vol. 8, no. 3, pp. 275-300.

[21] K. Yamanishi and J. Takeuchi 2001 "Discovering Outlier Filtering Rules from Unlabeled Data," Proc. Fourth Workshop Knowledge Discovery and Data Mining, pp. 389-394.

[22] K. Yamanishi and J. Takeuchi 2002 "A Unifying Approach to Detecting Outliers and Change-Points from Nonstationary Data," Proc of the Eighth ACM SIGKDD Int'l Conf. Knowledge Discovery and Data Mining.

[23] B.K. Yi, N.D. Sidiropoulos, T. Johnson, H.V. Jagadish, C. Faloutsos, and A. Biliris 2000 "Online Data Mining for Co-Evolving Time Sequences" Proc. 16th Int'l Conf. Data Eng.