International Journal of Recent Technology and Engineering (IJRTE)
ISSN: 2277-3878 (Online), Volume-8 Issue-6, March 2020

Machine Learning Techniques in RFID Datasets

Meghna Sharma, Manjeet Singh Tomer, Priyanka Vashisht

Abstract: ***In today's world of wireless communication, technologies like Wi-Fi, Global Positioning Systems (GPS), Radio Frequency Identification (RFID) and Wireless Sensor Networks (WSN) play a vital role in various applications for the benefit and convenience of the society. A widespread application of RFID networks based on automatic data capture technology is supply chain management. In RFID enabled supply chain process lot of possibilities of outliers or anomaly generation due to technical or environmental factors exist. This work mainly focuses on identifying various techniques used for outlier detection in RFID datasets in supply chain process. Most of literature studies are related to objects tracking and product management in the domain of supply chain but very few researchers have worked on the abnormal condition or outlier detection while monitoring of RFID tagged objects. Outliers are any kind of deviation in supply chain processes from its normal processing or behavior. Our research is specific to the supply-chain process using RFID system specifically for the abnormality detection in the localization process in supply-chain process. Inaccurate localization of objects can be due to several reasons like theft, counterfeiting, traffic problem, environmental factors or malfunctioning of the vehicle carrying RFID tagged objects.***

Keywords: RFID, Machine Learning, Wireless Communication, Outlier Detection

I. INTRODUCTION

Radio Frequency Identification (RFID) is still in the nascent stage due to challenges in shifting from the traditional approach which involves high cost of implementation and skills and time required in restructuring. The main reasons for the issues are lack of literature availability for the complete makeover from the non-RFID system to RFID implementation and its deployment in the current business processes. Improvements in supply chain were hindered due to lack of academic research and understanding of the technology though it's now-a-days taking up leap. We hope that this study can further develop insight into the challenges and opportunities of RFID and can direct academicians for further research on the areas of RFID that are most pertinent to practitioners. This paper provides the literature survey about techniques related to RFID data cleaning, outlier detection/path deviation techniques used in the domain of RFID-enabled supply chain and predictive analysis of outlier paths/trajectories.

Manuscript received on February 10, 2020.
Revised Manuscript received on February 20, 2020.
Manuscript published on March 30, 2020.
* Correspondence Author
Meghna Sharma*, Dept. of Computer Science and Engineering The NorthCap University, Gurugram meghnasharma@ncuindia.edu
Manjeet Singh Tomer , Dept. of Computer Applications J C Bose University of Science and Technology, YMCA, Faridabad
Priyanka Vashisht, Dept. of Computer Science and Engineering The NorthCap University, Gurugram priyankavashisht@ncuindia.edu

II. BACKGROUND

RFID is used to keep the track of processes and trace the supplies in construction and assembly industries. Ngai et al. [2] studied the literature of RFID technology by organizing their studies into technological based, application based, policy based and security-based categories. Not much of the literature was available to study the technology. Overall only 85 research papers were available. Their analysis gave useful insights on the anatomical details of the RFID literature. As RFID technology matured, so many applications [3] were unleashed to exploit inexpensive and highly available automatic identification. A complete framework for monitoring the progress using smart objects like RFID along with web service technologies in ubiquitous manufacturing had been proposed by Qu et al. [4].

RFID technology has shown the remarkable improvement in the domain of production planning and scheduling [5]. RFID systems do real time coordination and interaction among various levels like production level, planning level and scheduling level for achieving the lean control of processes in manufacturing [4].Smart manufacturing shop floors are created with RFID technology [6]. Supply chain with RFID technology has lot of advantages [7].

Some works focus on managing and mining RFID stream data. Hector Gonzalez et al. have done extensive work in this domain in various aspects. Traditional data warehousing multidimensional models won't fit into RFID datasets due to various properties of RFID data. A new model for warehousing RFID data has been proposed [8] in which object transitions' preservation is considered owing to the temporal feature RFID data. In order to represent the transportation of objects FlowGraph method has been proposed [9]. This representation can be very effective in the multi-dimensional analysis of flow of objects. Easy capturing of the movement of objects and monitoring exceptions in RFID flows has been proposed by using compressed probabilistic workflow method [10].

Elio Masciari [11] has researched on the outlier mining in RFID data stream to find out the outliers/anomalous nodes in the supply chain network. This research work used discrete Fourier-transform method to check for the similarities among various paths in the supply chain. Some research is done in area of mining RFID data.
Rule-and-Motif-based

anomaly detection framework has been designed for anomaly detection in moving object [12] but it does not fit well for the data with too many rules or outlier conditions as it adds to the complexity.

Partition-and-detect framework [13] for outlier detection of trajectory or path followed by the moving object had been proposed in which each trajectory or path followed by the object is first split into parts and then checked for similarity using all the angles between the distances in trajectory paths. This approach used clustering based approach but still didn't consider the speed variations of the objects in trajectories. Our research work has solved this issue though the base is similar to this research work. A new trajectory classification method for classifying various trajectories in supply chain called TraClass which uses hierarchical region-based approach is proposed by the same authors who used partition-and-group -detect framework [14] but it is not scalable.

Sensor network has some typical traits like limitation of resources, easy deployment of sensors, multiple hops in the network, massive data generation, and low maintenance requirement. Data mining of such type of datasets need to take into account these different constraints in the environment. Under the constraints of computational/memory/power limitations, A framework based on probabilistic method for supervised learning has been proposed by Ghosh et al. [15] considering computational, memory related or power related constraints. The main issue with supervised approach is the availability of training data which may not be the case in all type of applications. Moreover, we need to see over fitting problem too. A Spatio-Temporal Sensor Graphs (STSG) modelling based approach for mining sensor data for finding anomaly patterns and centralized locations at each time interval, has also been proposed [16] but for very long sequences it may not be very effective. An adaptive mining framework which can adapt according to changes in data has also been proposed by Cook et al. [17].

Internet of Things (IoT) has lot of contribution in data mining of different related domains, our main focus is on the one of the very important rudiment of IoT i.e. RFID. As a completely new paradigm in the research area, RFID-related applications still lack sufficient models and theories for the application of machine learning or data mining techniques. In the next section, a detailed study of the existing works done in the respective areas of sub-processes of the research work is explained along with the summarization for the research gaps.

III. EXISTING WORK

Our literature survey is done on basis of possible sub-processes used for RFID data mining and hence the discussions on the work/existing work in these areas.

A. The Data Pre-Processing Layer

It is divided into time pre-processing and path pre-processing management. It cleans the RFID event data obtained from the EPC network and providing the upper layer with clean and useful data by removing the possible false positives, false negatives and duplicate values.

RFID data generated by reader as read from tag is quite unreliable and redundant due to many external technical as well as environmental factors. Many Methodologies are proposed in literature to improve the reliability of RFID data. Work is done on both the hardware and the software aspects in a RFID system with respect to cleaning of RFID data [18]. Middleware solutions which are software based, refers to the implementation of algorithms for the correction of data streams coming through readers before being passed to the final database saved in the system for further analysis. There are many RFID-middleware solutions [19] [20] which are based on simple filtering techniques using fixed temporal sliding window filter to remove false negatives and false positives from RFID data and in these applications the very important thing is setting up of the window size for sampling the data and it's a drawback too. In the dynamic environment, it's not trivial to set up the appropriate smoothing-window size. Here the data generated in the environment is the continuous stream of datasets. A balance needs to be maintained between the tasks of ensuring completeness for the readings because of system unreliability and also ensure full capturing of the dynamics of tag as it moves in and out of the detection range of RFID reader. If a window size is selected as large then although it ensures the completeness but the system is not able to efficiently detect transitions of tags from inside to outside the window or outside to inside the window. On the other hand, if a small window size is selected then the system is able to detect transitions but it cannot ensure completeness due to missed readings. So if the size of window is set as small then it is possibility that some tags may miss being read leading to generation of false negative errors in which the tag is mistakenly assumed to be absent while it is actually present and if the size of window is set to very large then it can lead to generation of false positive errors as due to interpolation of the readings of the tag read by reader some tags which have already exited the detection region are also considered to be present and their information in stored by the system. So, in the real-world scenario experimentally no particular single sized window can consistently perform correctly for finding correct tag reads. We have studied and used an adaptive window-based approach called as window sub range transition detection algorithm (WSTD) [21] which can perform very well even in harsh environment with many dynamic changes. This approach also overcomes all the disadvantages of the fixed window protocol for filtering RFID data. The decision to use WSTD is based on the following literature study of the various techniques being used for the existing scenario.

Bai, Y, F Wang and P Liu [22] suggested the cleaning of RFID data in raw form into semantic application data. The false positive and negative readings as well duplicated readings should be filtered before being converted into the semantic form so that they can be used in different applications. The authors have proposed several effective methods to filter RFID data, for removing noise and eliminating duplicates.

They have used sliding window protocol with fixed size and threshold limit for removal of noise to be passed as input. In a dynamic environment of RFID data streaming continuously, setting up the parameters is quite a difficult task.

Gonzalez, H., J. Han and X Shen [23] proposed method based on Bayesian Networks which has the advantage of adjusting dynamically based on the probability of the tag existence. Due to its dynamism it is called as dynamic Bayesian network this method is computationally expensive due to sigma functions and cross-corpora calculations and performs very poorly in case the data set used is small.

Shen, H. and Y. Zhang [24] have used counting bloom filter-based technique known as decaying bloom filter. As new tags continuously enter in the sliding windows, the old and unused tags are removed. The method used can efficiently detect duplicate readings also. The problem with this filter is detection of false positive errors only but it cannot handle false negative errors.

Fan, H., Q.Y. Wu and Y.S. Lin. [25] proposed stream data processing by converting the semantic data in different logic rules to monitor the abnormal conditions of the work pieces in the manufacturing workshops. Work piece can be analyzed based on real time processing as well as history-oriented tracking. It works fine for small amount of data but doesn't fit well on large number of datasets.

Jeffery, Shawn R., Minos Greatlakes, and Michael J. Franklin [26] proposed an adaptive smoothing filter which can help in fixing the problems related to fixed window sliding protocol to reduce false positive and false negatives. An adaptive smoothing filter aggregates the RFID data and interpolates for lost readings. The algorithm known as Statistical sMoothing for Unreliable RFid data (SMURF) uses sampling theory and cleans the data by taking statistical sample of tag ids of the physical world and thus helps in modelling the unreliability of the RFID data readings by the reader. It uses binomial sampling and π-estimators; SMURF does the setting of correct window size automatically with continuous adaption over the historical and currently observed data readings. The effect of this cleaning algorithm is optimal only when the RFID tags move at a uniform speed. But in case of high-speed movement of tags in and out of readers' detection range the performance of this algorithm starts decreasing. Many variations like VSMURF [27] have been proposed based on SMURF concept.

Y. Wang, B.Y. Song, H. Fu, and X.G. Li [28] proposed a cleaning method KAL-RFID which is based on Kalman Filter. It is used to find false negative and false positive readings as well as solved the problem of delay occurrence in the transition time of tag stream. Kalman Filter update process consists of updating of time and measurement. This process needs lot of memory for storing the tags.

Wang, Y. L., C. Wang and X. H. Jiang [29] proposed a cleaning method based on bloom filter for handling the redundant data generated in the distributed data flow environment.

Massawe, L.V., J.D.M. Kinyua and H. Vermaak [30] proposed an improvement over SMURF which is also an adaptive sliding-window based approach known as Window Sub Range Transition Detection (WSTD). It can handle environmental variation and tag dynamics very efficiently. It can very well adjust the smoothing window size and thus can cope up with the changes in the environment which could lead to variations in the tag-reader performance. This thesis work cleaning method is based on WSTD for supply chain datasets. The existing work and the research gaps are summarized as shown in the Table I.

B. The Outlier Detection and Analysis Layer

It is the core of the three-layered system and it uses clustering-based outlier detection test in this work. Density based clustering technique is used for finding outlier clusters. Detection of outliers in RFID enabled supply-chain process requires the study on the trajectories identification which are not normal or deviating from the normal path called as outliers and to check for the similarity between various trajectories, different similarity measures are studied. The dataset formed in the supply-chain process is read in terms of trajectory as it contains the component of time and location in a particular order as read by the readers at various locations of the supply-chain path. In literature, several authors have used many types of techniques [30] like clustering [32], classification, and sequence pattern matching, probabilistic statistical techniques to find out outlier points or set of points from the given set of points. For the clustering approaches the training/labelled dataset is not required [33] but for the classification [31] there should be the availability of training points which is not available in all kind of applications. Any clustering algorithm/technique is based on the concept of finding dissimilarity/similarity between the objects to be clustered. The type of objects depends on the application being studied. Our research work deals with the trajectories followed by objects in the supply-chain path so here the similarities or differences between trajectories are considered and related work is studied. There are many similarity measures [34] studied and implemented with each having their own pros and cons. Euclidean distance [35] is the most commonly used distance measure with the condition of equal length in case of trajectory data [34] but it's not suitable for the cases in which the length of the trajectories are unequal. Also, it does not consider the time lagging factor within the path from source to destination. Other similarity measures like Hausdorff measure [36], Edit Distance [37], Fréchet Distance [38], Longest Common Subsequence [39], Dynamic Time Warping [40] etc. are also proposed for different applications. The following section covers the detail of related works in the area of trajectory mining as this is the base taken in our research for finding outlier points/nodes in the RFID enabled supply-chain network. Wang, Haozhou, et al. [34] has done a comparison of various trajectory similarity measures. Methods from time series analysis can be applied for the computation of trajectory similarity as the structure is same. Methods like Dynamic time warping (DTW), Longest Common Sub sequences (LCSS) and Edit Distance are quite commonly used methods. DTW, Edit Distance, and LCSS allow flexibility in finding match without any requirement of matching points at corresponding times.

DTW and Fréchet distance measures don't require exact time correspondence [54]. Time correspondence can't be ignored so simple Euclidian distance measure can't solve the purpose.

Berndt, D.J. and J. Clifford [40] proposed Dynamic Time Warping (DTW) distance measure as the technique to find the similarity between various speech patterns. Later this measure is used in the variety of problems in various other domains too. It is based on Euclidian distance but with the consideration of lagging or leading time factor over the complete path and thus is a solution to the weaknesses of Euclidian distance metrics. Due to this approach the time series or sequences which are not in phase locally due to temporal factor but are still similar are also taken into consideration. Although the time complexity of this approach is quadratic it is still considered an efficient way in terms of accuracy of finding similar time series/sequence data and is popularly used in various application areas like bioinformatics, medicine, engineering, entertainment etc.

Table-I: Summary Of Existing Approaches In RFID Data Cleaning

Sr. No.	Approach	Methodology	Research Gaps
1.	Bai, Y, F Wang and P Liu [22]	Fixed Window Protocol	Problem in setting window size
2.	Gonzalez, H., J. Han and X Shen [23]	Bayesian Networks	Computationally extensive, perform poorly for small datasets
3.	Shen, H. and Y. Zhang [24]	Decaying Bloom Filter	Can't detect false negative errors
4.	Fan, H., Q.Y. Wu and Y.S. Lin. [25]	Conversion of semantic data into logical rules	Not efficient for real time data processing
5.	Jeffery, Shawn R., Minos Garofalakis, and Michael J. Franklin [26]	Adaptive window sliding protocol	In case of high-speed movement of tags in and out of readers' detection range the performance of this algorithm starts decreasing.
6.	Y. Wang, B., Y. Song, H. Fu, and X., G. Li [28]	Kalman Filter	Process needs lot of memory for storing the tags.
7.	Wang, Y. L., C. Wang and X. H. Jiang. [29]	Based on Bloom Filter	Preprocess redundant data only, not false positives and false negatives

Jeung, H, et al. [42] proposed a hybrid prediction model to study the trajectory pattern being followed so that it can help in estimating status of any node in the trajectory network. Oobject's movements are based on many environmental factors like traffic jam and connected routes on road for vehicles, places of turbulence for aircraft, which makes use of mathematical formulas to represent the patterns followed in a path/trajectory, an inefficient way. So, the authors have proposed a novel approach which is a combination of Apriori [43] [44] for detecting frequent trajectory patterns and DBSCAN [45] for further clustering the sub trajectories. Use of Apriori for trajectory patterns however is not very memory efficient due to lot of candidates' generation in the intermediate steps.

Chun-Hee, L. and C Chin-Wan [46] proposed a path encoding schema using the concept of Chinese remainder theorem for the processing of large amount of RFID data for supply-chain management. Due to increasing number of tag numbers in system, the cost of storage of data and the time of processing are not utilized in an efficient manner.

Masciari, E. [47] proposed a system called as SMART (Simple Monitoring enterprise Activities by RFID Tags) which is based on defining a template for detection of outliers. The templates cover all the outlier scenarios and based on them the matching is done to find the outliers in RFID data streams. The templates consider sample taken(P), type of monitored objects (O) and the attributes (A), the outlier definition by means of a suitable function F (P, A, O) → {0, 1}. It is defined like a rule but for too many samples and objects this is not going to work efficiently. Scalability is an issue here.

Fan, H., et al. [25] proposed a model using the concept of a tree based structure path spliiting so that it movement of the products/tagged objects in trjactory path. This tree model finds out any deviation from the normal path and thus can be used to find out outliers, but for longer paths, it would increase the time and space complexity. Also redesigning of relational schema which can also store path and time information is required.

Hanning, C., et al. [48] proposed a novel method based on K means clustering algorithm [49]. Sequence of locations and time i.e. spatio-temporal elements are used to construct path network. Both K Means and Mean Shift [50] algorithms are compared for clustering similar paths and comparatively Mean Shift algorithm performed better but these algorithms need the number of clusters to be formed in advance and also there is the restriction of the spherical shape of the clusters.

Liu, X., et al. [51] worked on finding outliers in RFID trajectories by doing spatial analysis of the association among various discrete points in the path. Kriging method [52] is used for the interpolation of the number of points. Spatial and temporal variation in the accuracy of RFID readings is assessed quantitatively for finding or predicting the missed information values. It is not effective with large number of discrete points.

Kwon K., Kang D., Yoon Y., Sohn J.S., Chung I.J. [53] proposed a method called Procedure Tree for the mining of the massive data flow generated by tagged objects read by readers. The proposed system can perform better as compared to traditional systems for the tracking of objects but for longer supply-chain paths its efficiency decreases in terms of time and space utilization. Huang S.P., Wang D. [66] proposed distance based and rule-based approaches to detect the anomalies like delay in transiting and steal of the packages in the supply-chain path.

The system can provide some help for the enterprise management so as to help enterprises to effectively control the information of supply chain. The existing work and the research gaps are summarized as shown in the Table II.

Table-II: Summary Of Existing Approaches In RFID Outlier Detection Using Clustering

Sr. No.	Approach	Methodology	Research Gaps
1.	Wang, Haozhou, et al. [34]	Euclidian distance, Dynamic Time Warping measure, Longest Common Sub Sequences and Edit Distance measure are compared.	Euclidian Distance not appropriate approach with temporal factor
2.	Berndt, D.J. and J. Clifford [40]	Dynamic Time Warping as distance measure.	Quadratic Complexity but a good approach for comparison of paths with time and speed lags
3.	Jeung, H, et al [42]	Hybrid model (Apriori + DBSCAN)	Not memory efficient
4.	Chun-Hee, L. and C Chin-Wan [46]	Path Encoding Scheme	Memory and Time inefficient with increasing tag numbers
5.	Masciari, E. [47]	Rules/template creation	It doesn't work with high scalability
6.	Fan, H., et al. [25]	Tree based model	It doesn't work efficiently with long sequences of trajectory path.
7.	Huang, S.P. and D. Wang [66]	K Means and Mean Shift based Algorithm	Forms only spherical clusters and not efficient for finding outliers
8.	Liu, X., et al. [61]	Kriging Method of Interpolation	It doesn't work efficiently for dynamic system
9.	Kwon K., Kang D., Yoon Y., Sohn J.S., Chung I.J. [53]	Procedure Tree Method	It doesn't perform efficiently for longer sequences

C. PREDICTIVE ANALYSIS OF OUTLIER NODES

Predictive analysis of outlier nodes is done by recurrent-based neural networks, specifically Long Short-Term Memory (LSTM) [18]. The concept used in this research is based on the previous information about the location e.g. status of any particular node or set of nodes in RFID network paths. It can be an outlier point or non-outlier point. Location-based prediction is studied here. In our case the data is generated by applying TRAJODBSCAN, as proposed by us. Finally, outlier status is predicted based on the prediction of the next location. Lot of location-prediction based algorithms are available for prediction of next location if current location is known. The simplest way to handle this is by using speed and direction of movement but it's not easy in the real-world scenario in which the problems like traffic jam, theft, poor weather conditions do exist. As reviewed by Giannotti et al. [54] there are many methods for various applications specifically for data mining of trajectory. If the previous history data is available about the paths or trajectories being followed along with information about the deviation or outlier points in the path, prediction of any deviation or outlier path can be predicted. Main focus is improving the accuracy of prediction and according to literature available many techniques like pattern mining of moving objects [55] and model-based mining [56] are most commonly used. Li et al. has worked on the path prediction of [57] based on their moving pattern and behavior. The trajectory data is transformed in form of cell points for all the points and mining is done on that format. Yavas et al. [58] also worked on frequent pattern mining for path detection or deviation with Apriori algorithm as a base algorithm.

Locations, which are co-occurring, are extracted from the frequent patterns generated and analyzed. Further Morzy et al. also considered temporal and spatial features and developed a modified

Prefix-Span algorithm [59]. Gaussian model with mutli -centric approach is proposed by Cheng et al. [60] for prediction of similarity between different patterns of paths. The problem with this approach is no consideration of ordering of sequence. A hybrid technique which uses the Hidden Markov Model as base methodology has been proposed by Mathew et al. [61]. Jeung [42] used cell partition-based algorithm to map the trajectory points into frequent regions.Cell partition based method depends upon the granularity of cells for accuracy of prediction..

Sequence mining gained popularity and use of neural networks became the favorites for many researchers. Recurrent Neural Networks (RNN) [62] is very popularly used for time series data in wide areas of applications in the sequence mining [63]. In recurrent neural networks each layer actually represents each time step in the series of timestamp values. They are first kind of networks with internal memory and due to this reason; they are the most suitable ones for sequential data where previous steps need to be memorized. Liu et al. [64] used modified RNN based approach with spatial and temporal contexts for the prediction of space and time variant data. For small sequences or trajectories, Recurrent Neural Networks are good enough for time series data but not very promising when series or sequence is much longer say more than ten steps or hops. One of the biggest issue is vanishing as well as exploding gradients.

So many variations of Recurrent Neural Networks are available to overcome this big issue. We hypothesize that Long Short-Term Memory (LSTM) [18] which is an extension to Recurrent Neural Network discussed in the previous section may resolve the problem of handling bigger sequences or series .

LSTM architecture has memory blocks using different gates and connections for storing and memorizing the previous much older context and value also along with recent history of the sequence. Specifically, LSTM-based architecture is used for our outlier point's prediction in the supply-chain path. This concept has been used in this domain for the first time as per our knowledge.

Yavas G. Katsaros, D. Ulusoy O. Manolopoulos, Y. [58] proposed a three-layered framework for prediction of next point of users on personal communication network which is divided in the form of cells. In the first layer, the mobility patterns of the user are processed on the basis of the historical data of user trajectories. In the second layer, patterns are extracted in form of rules with constraints' consideration and mobility of users is considered and in the third and final layer; the rules generated in the previous layers are used to predict the path in the communication network for the users. The mobility rule-based prediction method is also compared with mobility prediction. Mobility prediction uses a Transition Matrix (TM) for saving the historical data and it is compared with ignorant prediction method which takes on the historical data and Ignorant Prediction method considers recent data. The accuracy of the proposed method is better but it takes toll on the memory requirements of the process.

Morzy M. [59] proposed movement rules method for finding frequent patterns in trajectories. Any trajectory followed by a moving object is compared against the saved movement rules and a probabilistic model to locate the objects in a path/trajectory is used. Proposed algorithm gives reasonably good prediction accuracy (80%). With the increasing network of trajectories, the proposed system can become very complex and with too many rules generation, the memory requirements would also be very high. Also matching with the rules will require higher processing time.

Cheng C., Yang H. King, I., Lyu M.R. [60] proposed a model based on matrix factorization method to find out the probability of check in by the user on any location in the path of the network. A framework with matrix factorization as well as social information of the user is used to demonstrate that the fused matrix factorization framework with multi Gaussian method uses the distance information and helps in the predicting the patterns of user check- ins in a particular network environment. Generation of matrix, though, is a high memory requirement process.

Jeung, H. Shen, H.T. Zhou, X. [42] proposed a novel approach which overcomes the problems associated with issues related to cell partitioning based processing. They have done detailed study based trajectory pattern models to find the association between the frequent regions and the partitioned cells using Hidden Markov process. With the proposed approach, the movement of any object is defined by the partitioned cells structure but the trajectory patterns used by the objects are defined by the frequent regions being followed by reading those cells. Deciding the granularity level of the cells is a problem. Moreover, this approach doesn't work well with longer trajectory paths.

Mathew W., Raposo R., Martins B. [61] proposed a representation learning method based on Hidden Markov Model (HMM) approach for Location-Based Social Networks, to be used for location recommendation and link prediction. The method helped in removing the drawbacks of the existing methods which focus only on topology patterns but not the sequences of check-ins. So, the approach works well in dynamic environment with hierarchical network. This approach however doesn't work well for longer sequences in a trajectory or path.

Kim, Moon-Chan, et al. [64] proposed a fuzzy cognitive map model. The weight matrix uses genetic algorithm with previous states data based on which the analysis of next state is done. It also took care of sudden change in the state and the cause for it base on the previous state data. The problem with this approach is that it does perform efficiently for very long paths or trajectory.

Graves, A. Mohamed, G. Hinton [65] studied about Recurrent Neural Networks (RNNs) for sequential data. The authors investigate deep recurrent neural networks and concluded RNN performance degrades with the increase in sequence path length. For longer path sequences Long Short-Term Memory Network which is a variant of RNN performs better. The existing work and the research gaps are summarized as shown in the Table III.

IV. LIMITATIONS OF THE EXISTING RESEARCH

There are certain limitations in the existing work specific to the analysis and data mining on RFID enabled supply chain processes, not much of literature is available. The related works are studied for similar kind of applications and data generated in RFID stream like traffic trajectories, healthcare applications, human trajectories etc. Here trajectory refers to the node's points combined to form a supply-chain path. The following conclusions are drawn after going through the complete study of existing literature on sub processes used in our research work i.e. data cleaning, outlier detection using clustering approach and predictive analysis of path points in a trajectory. Most of the existing work in data cleaning uses fixed window sliding protocol due to which there is a trivial problem of fixing up the window size for cleaning the false positives, false negatives and duplication in reading the tag data by RFID readers. Setting very small window size can result in generation of false negatives and setting up a very large window size can result in generation of false positives. Some techniques like Kalman filter and bloom filter need too much of memory and speed of processing is low. Adaptive window sliding detection is the best as it is dynamic window adjustment according to the stream of RFID tag data.

Most of works done for anomaly /outlier detection is based on clustering techniques. The main reason for it is non-availability of training data. Among the normal data the anomalous node point data is very less.

Retrieval Number: F9052038620/2020©BEIESP
DOI: 10.35940/ijrte.F9052.038620
Journal Website: www.ijrte.org

Published By:
Blue Eyes Intelligence Engineering
& Sciences Publication

The distance measures generally used are Euclidian, Fréchet, edit distance, longest common sub sequences, especially for trajectories but they have the drawbacks of either not taking into account the temporal factor into account with lag or lead of the objects due to speed variations or doesn't work effectively for longer trajectories.

Dynamic Time Warping is another similarity /distance measure which has a drawback of high complexity but works efficiently in case of trajectories of different lengths as well as time lags due to speed variations in the objects following the trajectories. To cluster the trajectories along with planned trajectories many clustering approaches like k means, mean shift, hierarchical clustering, cell-based partitioning is used but they either are not memory efficient, doesn't work well with longer trajectories or can't cope up with the dynamic and uncertain RFID data stream. Further for the sub process of predicting outliers using the previous history of trajectory data, many techniques like rules based, matrix based, hidden Markov model based, decision tree based, neural networks based are used but they don't work efficiently in case of long length trajectories.

Table-III: Summary Of Existing Approaches In RFID Outlier Detection Using Classification

Sr. No.	Approach	Methodology	Research Gaps
1.	Yavas G. Katsaros, D. Ulusoy O. Manolopoulos, Y. [58]	Rules based	Memory inefficient and high time processing requirements
2.	Morzy M. [59]	Rules based	Memory inefficient and high time processing requirements
3.	Cheng C., Yang H. King, I., Lyu M.R. [60]	Factorization matrix	Memory inefficient
4.	Jeung, H. Shen, H.T. Zhou, X. [42]	Hidden Markov Model based	It doesn't work efficiently with very long trajectory paths
5.	Mathew W., Raposo R., Martins B. [61]	Hidden Markov Model based	It doesn't work efficiently with very long trajectory paths
6.	Kim, Moon-Chan, et al [64]	Fuzzy Cognitive Networks with Genetic Algorithm	It doesn't work efficiently with very long trajectory paths

V. Conclusion and Future Scope of The Research

The literature study gives the insight for the development of a complete framework to find out the outliers in the RFID enabled supply-chain path. As per the understanding and study till now there is no availability of such system in the mention domain. A system which can find out outliers with good accuracy with the consideration of long path sequences of the trajectories followed by the objects with varying speed and acceleration is required. The study conducted in this paper aims for the same and to outline the so that the researchers can overcome the problems in the existing systems. Scope of research is quite vast as the framework designed can be used to find out outliers or anomalies in various applications which are RFID enabled. Apart from outlier detection of RFID supply chain datasets, they can also be used in RFID path deviation detection, RFID enabled healthcare process, RFID enabled toll process, RFID enabled car parking, RFID enabled tracking of things and so on.

References

1. Y. J. Shin, J. S. Oh, S. H. Shin and H. L. Jang, "A Study on the Countermeasures of Shipping and Port Logistics Industry in Responding to the Progression of Fourth Industrial Revolution," Journal of Korean Navigation and Port Research, vol. 42, no. 5, pp. 347-355, 2018.
2. E. W. T. Ngai, K. K. Moon, F. J. Riggins and Y. Y. Candace , "RFID research: An academic literature review (1995–2005) and future research directions," International Journal of Production Economics, vol. 112, no. 2, pp. 510-520, 2008
3. B. Nath, F. Reynolds and R. Want, "RFID Technology and Applications," IEEE Pervasive Computing, vol. 1, pp. 22-24, 2006.
4. T. Qu, H. D. Yang, G. Q. Huang, Y. F. Zhang, H. Luo and W. Qin, "A case of implementing RFID-based real-time shop-floor material management for household electrical appliance manufacturers.," Journal of Intelligent Manufacturing, vol. 23, no. 6, pp. 2343-2356, 2012.
5. R. Zhong, G. Huang, S. Lan, Q. Dai, X. Chen and T. Zhang, "A big data approach for logistics trajectory discovery from RFID-enabled production data," International Journal of Production Economy, vol. 165, p. 260–272, 2015.
6. Zhong, Ray Y., "Analysis of RFID datasets for smart manufacturing shop floors." in 15th International Conference on Networking, Sensing and Control (ICNSC). IEEE, 2018, pp. 1-4.
7. A. Sarac, N. Absi and S. Dauzère-Pérès, "A literature review on the impact of RFID technologies on supply chain management," International Journal of Production Economics, vol. 128, no. 1, pp. 77-95, 2010.
8. J. Han, H. Gonzalez, X. Li and D. Klabjan, "Warehousing and mining massive RFID data sets," in International Conference on Advanced Data Mining and Applications. Springer, Berlin, Heidelberg, 2006, pp. 1-18.
9. H. Gonzalez, J. Han and X. Li, "FlowCube: Constructuing RFID FlowCubes for Multi-Dimensional Analysis," in VLDB 2006, Seoul, Korea, 2006, pp. 834-845.
10. H. Gonzalez, J. Han and X. Li, "Mining compressed commodity workflows from massive RFID data sets," in Proceedings of the 15th ACM international conference on Information and knowledge management, 2006, pp. 162-171.
11. E. Masciari., "A Framework for Outlier Mining in RFID data," In: 11th International Database Engineering and Applications Symposium (IDEAS 2007), IEEE, 2007, pp. 263-267.
12. X. Li, J. Han, S. Kim and H. Gonzalez, "ROAM: Rule- and Motif-Based Anomaly Detection in Massive Moving Objects Dataset," in Proceedings of the 2007 SIAM International Conference on Data Mining. Society for Industrial and Applied Mathematics, 2007, pp. 273-284.
13. J.-G. Lee, J. Han and K.-Y. Whang, "Trajectory clustering: a partition-and-group framework," in Proceedings of the 2007 ACM SIGMOD international conference on Management of data, 2007, pp. 593-604.
14. J.-G. Lee, J. Han, X. Li and H. Gonzalez, "TraClass:trajectory classification using hierarchical region-based and trajectory-based clustering," in Proceedings of the VLDB Endowment, vol. 1, no. 1, pp. 1081-1094, 2008.

15. J. Ghosh, "A Probabilistic Framework for Mining Distributed Sensory Data under Data Sharing Constraints," in International wokshop on Knowledge Discovery from Sensors, vol. 7, no. 1, pp. 1-7, 2008.
16. B. George, J. M. Kang and S. Shekhar, "Spatio-Temporal Sensor Graphs (STSG): A data model for the discovery of spatio temporal patterns," Intelligent Data Analysis, vol. 13, no. 3, pp. 457-475, 2009.
17. P. Rashidi and D. J. Cook, "An Adaptive Sensor Mining Framework for Pervasive Computing Applications," in International Workshop on Knowledge Discovery from Sensor Data. Springer, Berlin, Heidelberg, 2008, pp. 154-174.
18. P. Malhotra, L. Vig, G. Shroff and P. Agarwal, "Long short term memory networks for anomaly detection in time series.," in Proceedings Presses universitaires de Louvain, vol. 89, pp. 89-93, 2015.
19. S. Lv and S. A. Yu, "Middleware-Based Algorithm for Redundant Reader Elimination in RFID Systems," ACTA ELECTRONICA SINICA, vol. 40, no. 5, pp. 965-970, 2012.
20. H. Ziekow, L. Ivantysynova and O. Günter, "RFID Data Cleaning for Shop Floor Applications," in Unique Radio Innovation for the 21st Century, Berlin, Heidelberg, Springer, 2011, pp. 143-160.
21. H. Xu, J. Ding, P. Li and L. Wei, "A Review on Data Cleaning Technology for RFID Network.", in International Conference on P2P, Parallel, Grid, Cloud and Internet Computing. Springer, Cham, 2016, pp. 373-382.
22. Y. Bai, F. Wang and L. Peiya, "Efficiently filtering RFID data streams," in Proceedings of the CleanDB Workshop, Seoul, Korea, September 2006.
23. H. Gonzalez, J. Han and X. Shen, "Cost-Conscious Cleaning of Massive RFID Data Sets," in 23rd International Conference on Data Engineering. IEEE, 2007, pp. 1268-1272.
24. H. Shen and Y. Zhang, "Improved approximate detection of duplicates for data streams over sliding," Journal Computer Science Technology, vol. 23, pp. 973–987, 2008.
25. H. Fan, Q. Wu, Y. Lin and J. Zhang, "A split-path schema-based RFID data storage model in supply chain management," Sensors , vol. 13, pp. 5757–5776, 2013.
26. S. R. Jeffery, M. Garofalakis and M. J. Franklin, "Adaptive cleaning for RFID data streams," in Proceedings of the 32nd international conference on Very large data bases. VLDB Endowment, 2006, pp. 163-174.
27. H. Xu, W. Shen, P. Li, D. Sgandurra and R. Wang, "VSMURF: A Novel Sliding Window Cleaning Algorithm for RFID Networks," Journal of Sensors, pp. 1-11, 2017.
28. Y. Wang, B.-Y. Song, H. Fu and X.-G. Li, "Cleaning method of RFID stream based on Kalman filter," Journal of Chinese Computer Systems, vol. 32, no. 9, pp. 1794–1799, 2011.
29. Y. L. Wang, C. Wang and X. H. Jiang, "RFID duplicate removing algorithm based on temporal-spatial Bloom Filter," Journal of Nanjing University of Science and Technology, vol. 39, no. 3, pp. 253-259, 2015.
30. L. Massawe, J. Kinyua and H. Vermaak, "Reducing false negative reads in RFID data streams using an adaptive sliding-window approach," Sensors, vol. 12, pp. 4187–4212, 2012.
31. Y. Zhao, "Data mining techniques," unpublished.
32. D. Cui and Q. Zhang, "The RFID data clustering algorithm for improving indoor network positioning based on LANDMARC technology," Cluster computing, vol. 22, pp. 5731-5738, 2019.
33. P.-N. Tan, M. Steinbach and V. Kumar, "Data mining cluster analysis: basic concepts and algorithms," Introduction to data mining, 2013, pp. 487-533.
34. H. Wang, "An effectiveness study on trajectory similarity measures," in Proceedings of the Twenty-Fourth Australasian Database Conference, Australian Computer Society, Inc., vol. 137, 2013, pp. 13-22.
35. S. Santini and R. Jain, "Similarity measures," IEEE Transactions on pattern analysis and machine Intelligence, vol. 21, no. 9, pp. 871-883, 1999.
36. W. J. Rucklidge, "Efficiently locating objects using the Hausdorff distance," International Journal of computer vision, vol. 24, no. 3, pp. 251-270, 1997.
37. E. S. Ristad and P. N. Yianilos, "Learning string-edit distance," IEEE Transactions on Pattern Analysis and Machine Intelligence, vol. 20, no. 5, pp. 522-532, 1998.
38. T. Eiter and H. Mannila, "Computing discrete Fréchet distance.," Tech. Report CD-TR 94/64, Information Systems Department, Technical University of Vienna, pp. 636-637, 1994.
39. M. Paterson and V. Dančík, "Longest common subsequences," in International Symposium on Mathematical Foundations of Computer Science, Springer, Berlin, Heidelberg, 1994, pp. 127-142.
40. D. J. Berndt and J. Clifford, "Using dynamic time warping to find patterns in time series," in KDD workshop. 1994, pp. 359-370.
41. K. Buchin, M. Buchin and C. Wenk, "Computing the Fréchet distance between simple polygons in polynomial time," in Proceedings of the twenty-second annual symposium on Computational geometry. 2006, pp. 80-87.
42. H. Jeung, Q. Liu, H. T. Shen and X. Zhou, "A hybrid prediction model for moving objects," in 24th international conference on data engineering, IEEE, 2008, pp. 70-79.
43. A. Monreale, F. Pinelli, R. Trasarti and F. Giannotti, "Wherenext: a location predictor on trajectory pattern mining," in Proceedings of the 15th ACM SIGKDD international conference on Knowledge discovery and data mining, 2009, pp. 637-646.
44. I. Akihiro, T. Washio and H. Motoda, "An apriori-based algorithm for mining frequent substructures from graph data," in European conference on principles of data mining and knowledge discovery. Springer, Berlin, Heidelberg, 2000, pp. 13-23.
45. J. Erman, M. Arlitt and A. Mahanti, "Traffic classification using clustering algorithms," in Proceedings of the 2006 SIGCOMM workshop on Mining network data, 2006, pp. 281-286.
46. C.-H. Lee and C.-W. Chung, "RFID data processing in supply chain management using a path encoding scheme," IEEE transactions on knowledge and data engineering, vol. 23, no. 5, pp. 742-758, 2011.
47. E. Masciari, "Smart: Stream monitoring enterprise activities by RFID tags," Information Sciences, 2012 - Elsevier, vol. 195, pp. 25–44, 2012.
48. C. Hanning, Y. Z and K. H, "Multi-colony foraging optimization with cell-to-cell communication for RFID network planning," Applied Soft Computing , vol. 10, no. 2, pp. 539-547, 2010.
49. J. A. Hartigan and M. A. Wong, "Algorithm AS 136: A k-means clustering algorithm," Journal of the Royal Statistical Society. Series C (Applied Statistics), vol. 28, no. 1, pp. 100-108, 1979.
50. Y. Cheng, "Mean shift, mode seeking, and clustering," IEEE transactions on pattern analysis and machine intelligence, vol. 17, no. 8, pp. 790-799, 1995.
51. X. Liu, J. Shannon, H. Voun, M. Truijens, H. Chi and X. Wang, "Spatial and temporal analysis on the distribution of active radio-frequency identification (RFID) tracking accuracy with the kriging method," Sensors, vol. 14, no. 11, pp. 20451-20467, 2014.
52. N. Cressie, "The origins of kriging," Mathematical geology, vol. 22, no. 3, pp. 239-252, 1990.
53. K. Kwon, D. Kang, Y. Yoon, J. Sohn and I. Chung, "A real time process management system using RFID data mining," Computers in Industry, vol. 65, p. 721–732, 2014.
54. F. Giannotti, M. Nanni, F. Pinelli, D. and Pedreschi "Trajectory pattern mining," in: Proceedings of the 13th ACM SIGKDD international conference on Knowledge discovery and data mining, 2007, pp. 330-339.
55. H. Cao, N. Mamoulis and D. W. Cheung, "Mining frequent spatio-temporal sequential patterns," in Fifth IEEE International Conference on Data Mining (ICDM'05), IEEE, 2005.
56. W. M. Van der Aalst, M. H. Schonenberg and M. Song, "Time prediction based on process mining," Information systems, vol. 36, no. 2, pp. 450-475, 2011.
57. Q. Li, Z. Zeng, T. Zhang, J. Li and Z. Wu, "Path-finding through flexible hierarchical road networks: An experiential approach using taxi trajectory data," International Journal of Applied Earth Observation and Geoinformation, vol. 13, no. 1, pp. 110-119, 2011.
58. G. Yavaş, D. Katsaros, Ö. Ulusoy and Y. Manolopoulos, "A data mining approach for location prediction in mobile environments," Data Knowledge Engineering , vol 54, no. 2, pp. 121-146, 2005.
59. M. Morzy, "Mining frequent trajectories of moving objects for location prediction," in International workshop on machine learning and data mining in pattern recognition. Springer, Berlin, Heidelberg, 2007, pp. 667-680.
60. C. Cheng, H. Yang, I. King and M. R. Lyu, "Fused matrix factorization with geographical and social influence in location-based social networks," in Proceedings of the AAAI Conference on Artificial Intelligence, 2012, pp. 17-23.
61. W. Mathew, R. Raposo and B. Martins, "Predicting future locations with Hidden Markov Models," In: Proceedings of the 2012 ACM conference on ubiquitous computing, 2012, pp. 911-918.
62. I. Sutskever, O. Vinyals and Q. V. Le, "Sequence to sequence learning with neural networks," in Advances in neural information processing systems, 2014, pp. 3104-3112.
63. C. C. Aggarwal, M. A. Bhuiyan and M. A. Hasan, "Frequent pattern mining algorithms: A survey," in Frequent pattern mining, Springer, Cham, 2014, pp. 19-64.

64. M.-C. Kim, C. O. Kim, S. R. Hong and I.-H. Kwon, "Forward–backward analysis of RFID-enabled supply chain using fuzzy cognitive map and genetic algorithm," Expert Systems with Applications, vol. 35, no. 3, pp. 1166-1176, 2008.
65. A. Graves, A. R. Mohamed and G. Hinton, "Speech recognition with deep recurrent neural networks," in international conference on acoustics, speech and signal processing, IEEE, 2013, pp. 6645-6649.
66. S. Wu, P. Bertholet, H. Huang, D. Cohen-Or, M. Gong, and M. Zwicker, M., " Structure-aware data consolidation," IEEE transactions on pattern analysis and machine intelligence, vol. 40, no. 10, pp. 2529-2537, 2017.

AUTHORS PROFILE

Dr Meghna Sharma is associated with NCU since 2008. She has an excellent teaching and research experience of more than 15 years. She is B. E in Computer Science and Engineering from CRSCE Murthal, M.Tech in Computer Science and Engineering from GJU Hissar and PhD from YMCA Faridabad. She has also completed a reputed PG certification of Big Data Engineering from BITS Pilani with UpGrad and did many projects related to Big Data Engineering. She has done certification courses from Coursera of R Programming and Getting and Cleaning Data by Johns Hopkins University and Pattern Discovery in Data Mining by University of Illinois.

Dr. Manjeet Singh is Professor and Proctor in Department of Information Technology and Computer Application at J C Bose University of Science and Technology, YMCA, Faridabad. He has completed his Ph.D in year 2008 from Maharshi Dayanand University and M.Tech in 2002 from Guru Jambeshwar University. His area of interest are Artificial Intelligence, Soft Computing, Natural Language Processing, Computer Networks, Ad-Hoc Networks, Information Retrieval, Semantic web, Compiler Design, and Data Structures and Algorithm Design.

Dr Priyanka Vashisht joined NCU in 2019 as Assistant Professor in Department of CSE. She has an excellent teaching experience of about 15 years in various esteemed institutions. She has earned her Ph.D from Thapar University, Patiala and M.Tech from Banasthali Vidyapeeth, Banasthali. Her current areas of research include Grid Computing, Cloud Computing, Distributed Systems and Internet of Things (IoT) and Big Data. She has guided more than 70 B.Tech projects. She has published various Papers in peer reviewed International Journals with good indexing and reputed national/international conference proceedings. She was member of Board of Studies at GGSIPU University. She is member of Computer Science Teaching Association (CSTA) and ACM.