Outlier Detection in RFID Datasets in Supply Chain Process: A Review

Meghna Sharma
Assistant Professor (Sr. Scale),
Department of CSE/IT,
ITMU, Gurgaon, Haryana

Manjeet Singh, PhD.
Associate Professor,
Department of CS
YMCA University, Faridabad, Haryana

ABSTRACT

Outlier detection has been a very important concept in the realm of data analysis. Most real-world databases include a certain amount of exceptional values, generally termed as “outliers”. The finding of outliers is important for improving the quality of original data and for reducing the impact of outlying values in the process of knowledge discovery in databases.. Outlier detection has been researched within various application domains and knowledge disciplines. Supply Chain Process is one of the popular and important domains. The implementation of RFID leads to improved visibility in supply chains. However, as a result of the increased collection of data and data granularity, new data management challenges are faced by supply chain participants new techniques for outlier detection are experimented. In this Paper the problem of detecting outliers in RFID readings stream. is addressed and considering the stream based ,spatio-temporal nature of RFID datasets, density based outlier detection technique is concluded to be the best among all the existing approaches. for outlier detection

General Terms

RFID, Outlier Detection

Keywords

Outlier Detection, RFID, Supply chain process Density Based, Data Mining.

1. INTRODUCTION

Introduction is divided into two major sections. First section describes about basics of Radio Frequency Identification, RFID in Supply Chain Process, and Architecture of RFID Warehouse. Next Section gives introduction to Outlier Detection and description of Outlier Detection in RFID Supply Chain Datasets.

1.1 RFID

Radio Frequency Identification (RFID) is a type of automatic identification system. An RFID system enables the transmission of data by a portable device, called a tag, which is read by a reader of RFID and processed according to the requirement of any specific application. The data transmitted by the tag provides identification or location information, or tagged product specifications, such as price, color, date of purchase, etc. The use of RFID in tracking and access applications first appeared during the 1980s. RFID quickly gained attention because it can track moving objects very quickly and that too in an automated ways.

In a typical RFID system, an inexpensive tag with a transponder with a digital memory chip recognized by a unique electronic product code, is put along with individual objects An antenna with a transceiver and decoder, emits a signal and read data from and write data in it, by activating the RFID . After passing of an RFID tag through the electromagnetic zone, the antenna or interrogator detects the reader's activation signal. The data encoded in the tag's integrated circuit (silicon chip) is decoded by the reader and is passed to the host computer for its further processing. Figure 1 gives the basic structure of RFID system.

The applications of Radio Frequency Identification (RFID) are one of the most emerging key components in object tracking and supply chain management systems. In future almost every major retailer will use RFID systems to track the shipment of products from suppliers to Warehouses. This data along with providing an insight into shipment and other supply chain process efficiencies is also capable of determination of product seasonality and other trends and provides very important information and analysis for the company’s plans. So many advanced uses of RFID are being explored in a wide range of applications. For example, tire manufacturers plan to embed RFID chips in tires to determine the tire deterioration. Many pharmaceutical companies are embedding RFID chips in drug containers to better track and avert the theft of highly controlled drugs. Airlines and satellites are considering RFID-enabling key onboard parts and supplies for the optimization of aircraft maintenance and turnaround time of airport gate preparation.

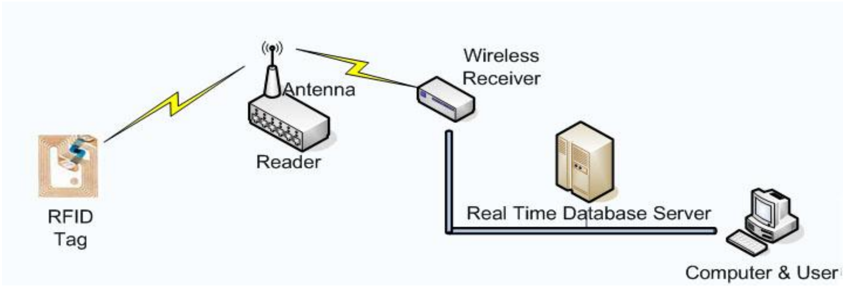

Figure 1: Basic RFID System

Some specific properties of RFID datasets which make them to be handled differently from the non- RFID datasets [1] are as follows:-

- **Temporal and Dynamic**: Applications dynamically generate observation (readings). Objects location as well as containment relationship among objects change along the time as a result, RFID data carry state changes.
- **Inaccuracy of Data:** Sometimes non-existing tag may be incorrectly read (False positive reads) or reader may miss tags which were in its vicinity (False negatives). Also reader may read a same tag more than once. Such an erroneous data should be filtered in a semantic manner.
- **Continuous Streaming:** Number of RFID tags are proportional to number of items being serviced/tracked and number of readers are proportional to traceable strategic locations/areas. In typical scenarios, tagged object stay in place for longer duration and readers records their existence on continuous basis in periodic intervals. A tuple is inserted into database each time a tag is read by reader. All these small observations pile up to produce redundant data.. This continuous streaming data must be filtered. Also some kind of compression is needed to make reduce data without loss of information.
- **Granularity:** The level of granularity for data collection needs to be determined. This factor depends on applications for which RFID system is being implemented e.g. the granularity of data collection can be a single item on a shelf or in store if we take retail store in a scenario. On the contrary, the granularity will be unit of luggage/baggage., if a n RFID system is deployed in an airport.

1.2 RFID in Supply Chain Process

The direct benefits of using RFID technology in supply chain management and retailers [2].

- ***Automation:*** The most direct benefits of RFID is an automatic version of the barcode. Using RFID can save a huge amount of labor because we do not need to use humans to scan the items in different parts of the organization.
- ***Inventory Shrinking:*** Retailers replenishment decisions are based on the inventory information kept in the inventory system which is assumed to be accurate. However, sometimes the count in the inventory system does not reflect the correct number of items in the actual inventory due to shrinkage or stock loss. Handling this type of problem is a very costly operation that requires a regular manual counting of stock. RFID technology will decrease the cost of inventory counting significantly and this process can be easily accomplished regularly.
- ***Inventory Replenishment*:** Shelf inventory management is critical to the retail business. Quite often, items are out of stock on the store shelf while there is still plenty of stock available in the backroom of the store. This is because there is no automatic process for detecting the stock out and restocking the shelf once it becomes empty. With RFID technology, shelf Inventory can be tracked automatically. For example, if a customer buys the item that makes the quantity move below the threshold available on the shelf, an automatic restock order will immediately be sent to the store manager for refilling.
- ***Visibility Of Inventory Across The Supply Chain:*** The RFID technology provides a comprehensive visibility of inventory throughout the entire supply chain. Each product item is tagged. So, items that moving around a business can be monitored the from supplier warehouse to the shelves. This comprehensive view of the supply chain helps in reducing the inventory, streamline logistics, and helps in optimizing the efficiency of their workforces as they gain a competitive advantage.

The flow in a supply chain process is explained in figure 2.

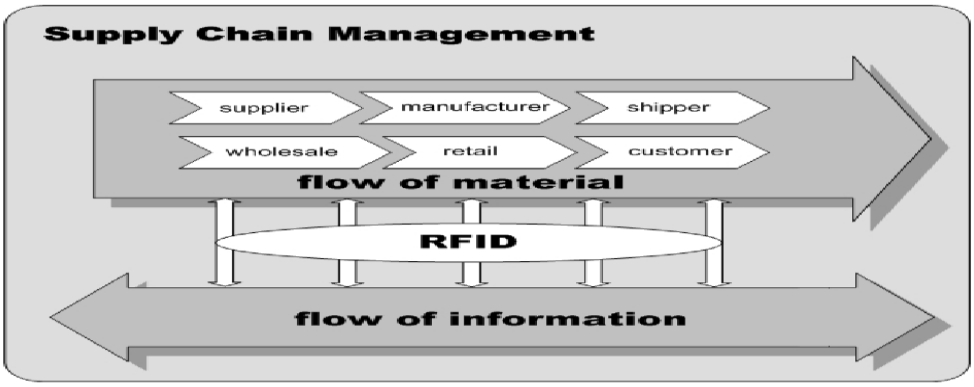

Figure 2: Flows in Supply Chain Process[3]

1.3 General RFID Data Management Challenges

Data capture by means of RFID requires no manual actions, which causes a decrease in the cost of the data capture process. This implies that the number of capture sessions and points increases. For instance, with the implementation of the so-called smart shelves, data from the objects on the shelves can be read continuously. This has multiple effects:

- The data quantity increases significantly.
- RFID data is quite dynamic, because of continuous readings.
- The data accumulates continuously (in an asynchronous manner). These data must be processed in real time and then relayed to the connected systems.
-

Special Warehousing is needed for such RFID data which is already been handled [4]

2. OUTLIER DETECTION

The term Outlier can generally be defined as an observation that is significantly different from the other values in a data set. The outliers are the instances of error or indication of the events. The main aim of Outlier Detection is to identify such outliers with the improvement in the analysis of data and further discovery of interesting and useful knowledge about unusual events within numerous applications domains.

Outliers are mostly the indication of an error or an event Outliers occur due to the following reasons [5].

***Error**:* - This sort of outliers is also known as anomalies, discordant observations, exceptions, faults, defects, aberrations, noise, damage or contaminants. This may occur because of human errors, instrument errors, mechanical faults or change in environment.

***Event**:* - Outliers may be generated by different mechanism, which indicates that this type of outliers belong to unexpected patterns that do not conform to normal behavior and may include useful and interesting information about rarely occurring events within the numerous application domains.

2.1 Outlier Detection in Supply Chain Process

Outlier detection in supply chain process can be explained with respect to localization, tracking, logistics and transportation. Localization refers .to the determination of location of objects or set of objects .Logistics refers to manage and control the flow of products from source to destination .Outliers in localization would be filtering any erroneous data due to inaccurate localization from the collection of raw data used to calibrate the nodes of network while simultaneously tracking the moving target Outliers in tracking and tracing shipments would be finding inappropriate quantity of product and giving notification to all trading partners in time for the same. Tracing shipments could find inappropriate quantity and quality of the product and notify all trading partners in time.

Due to the streaming nature of RFID readings, large amounts of data are generated by RFID devices. In particular, RFID applications will generate a lot of data pertaining to time and location. Every individual item can be tagged thus leaving a “trail” of data as it moves across different locations. This kind of data follows a particular sequence moving in a particular direction changing location and time. Trajectory data would be the terminology given to such kind of data.

This scenario raises new challenges in effectively and efficiently exploiting such large amounts of data with special features. Outlier Detection is a technique for detecting anomalous data in order to prevent problems related to inefficient shipments in any supply chain process, thus can control correct shipment of items [6]. There are various techniques available to do outlier detection with each having its own pros and cons. In the next section the survey for Outlier Detection techniques for their comparison with reference to the data type is discussed and challenges faced are being referred to.

3. CURRENT TRENDS

This section defines various Outlier Detection techniques currently used for various application domains, classification of various outlier detection techniques in tree structure, then explanation of existing techniques with the challenges associated with them. Then the Outlier Detection approach for the spatio- temporal continuously streaming dataset, which is specifically the dataset type of RFID dataset too, is discussed with current methodologies available for them as per survey.

Outlier Detection can be categorized based on availability of pre labeled data into three categories [7].

- ***Supervised Learning Approach***:- In this approach normality and abnormality models are learnt by using some pre-labelled data called test data and then classification of the new data point is done as normal and abnormal based on the basis of its fitting into either of the one model.
- ***Unsupervised Learning Approach***:-These approaches can identify outliers without the use of pre-labelled data. They are more general as compared to supervised learning approaches.
- ***Semi Supervised Learning Approach***:-These approaches can only require training on pre labelled normal data to learn boundary of normality and classify a new data point as normal and abnormal depending on how well data point fit into the normality model.

In Figure 3, complete classification of outlier detection techniques based on the current trends is shown.

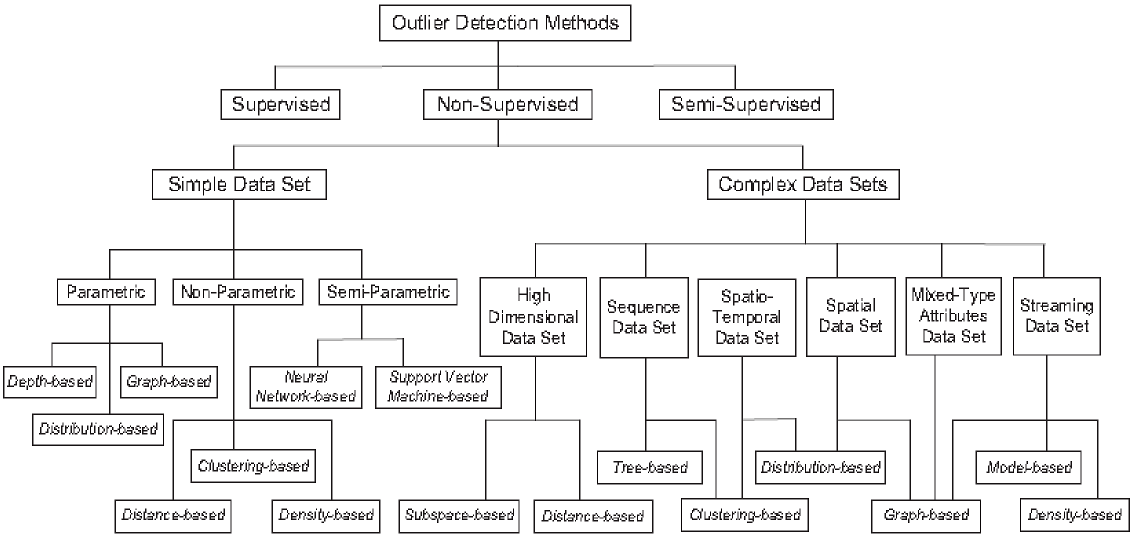

Figure 3: Outlier Detection Method Classification

Due to the uncertainty of RFID data with its trajectory characteristics, supervised and semi supervised approaches won’t fit in well for outlier detection .Best would be unsupervised one and based on further classification of unsupervised one for both simple data set and complex data sets, the approaches which can be used to handle RFID datasets which is spatio-temporal, streaming and sequence based are:

- ***Distribution-Based*** [8] methods assume that the whole data follow a *standard statistical* distribution model and determine a point as an outlier depending on whether the point deviates significantly from the data model. These methods can identify outliers on the basis of an appropriate probabilistic data model. in fast and efficient manner
- ***Graph-Based*** [9] methods make use of a powerful tool *data image* and then map the data into a *graph* so that visualization of the single or multi-dimensional data spaces can be done. Particular. Positions of the graph can be termed as Outliers.. These methods are good for the identification of outliers in real-valued and categorical data.
- ***Clustering-Based*** [10] *Earlier clustering based methods* can optimize the process of *clustering* of data, where outlier detection are only by-products of no interest. The new clustering-based outlier detection methods can identify outliers as points that do not belong to clusters of a data set or as clusters that are significantly smaller than other clusters very effectively
- ***Density-Based*** [11] methods take the local *density* into account when searching for outliers. These methods are suitable to identify local outliers in data sets with diverse clusters very efficiently.
- ***Tree-Based*** [12]methods do the construction of a specific tree as an index for the decomposition of data structure and use an efficient measure of similarity for the sequence data so that outliers can be distinguished from non outliers. These methods efficiently can segregate outliers by examining nodes near the root of tree.
- .***Model-Based*****[13]** methods find out outliers by the constructing a *model*, that can show the statistical behavior of a data stream. Those points that deviate

from the learned model are outliers. The streaming online data can be very efficiently These methods can efficiently dealt with these methods.

All the above mentioned approaches can be related to the outlier detection of RFID datasets. RFID datasets vary in structure and comparison among various techniques depends on the application used and the kind of outliers one is interested in. Datasets can be differentiated based on number of dimensions and data types .Data can be univariate and multivariate .Univariate data is collected for one variable only in each sample and multivariate data is the data collected for several variables in each sampling unit .RFID data even in raw form, consists of at least four variables for tag id, reader location, time in and time out..If it is processed for an application like supply chain management, raw data can be further mapped with more number of variables like granularity level, path flow level etc. RFID dataset then has to be multivariate .Depending on the parameters like technique used ,number of outliers detected, data dimension and data type, a comparison is done and analyzed for the best possible technique for outlier detection of RFID datasets. Table 1 shows the comparison of various technique and data types.

Table 1: Classification and Comparison of Outlier Detection Techniques for Complex Data

Technique Used	Number Of Outliers Detected At Once	Data Dimension	Data Type
Graph Based	One/Multiple	Multivariate Moderate	Mixed type, streams
Clustering Based	Multiple	Multivariate Moderate	Sequence
Tree Based	Multiple	Multivariate Moderate	Sequence
Distribution Based	One/Multiple	Univariate Multivariate moderate	Spatial
Model Based	Multiple	Multivariate Moderate	Streams
Density Based	Multiple	Multivariate Moderate	Streams

4. CONCLUSION

Model Based, Graph Based and Density Based can be appropriate for the Outlier Detection in RFID Datasets, but Model-Based approaches represent the statistical behavior of data stream by the construction of a model and further declare those points that deviate from this model as outliers. They can be used to efficiently deal with the streaming online data ; however, it is not easy to construct an accurate model to represent the whole data. Graph-based approaches identify outliers based on the estimation of the distribution of data stream. but, they may suffer from the curse of dimensionality and the accuracy of the estimation of data distribution. Density Based Approach is the best among all of these considering all the factors and type of RFID datasets. RFID datasets need a special mining as it is a combination of multiple types of datasets .It normally follows a particular trajectory which refers to a particular value of location at a particular and it keeps on changing with time continuously. Trajectory data mining [14] is a current trend to handle such kind of data though the techniques are same and density based is again considered optimal as per study done. Taking density based approach as a base, further work can be done for the outlier analysis of RFID datasets in supply chain process more efficiently.

5. REFERENCES

[1] Finkenzeller ,K.,Waddington,R., -eds :RFID Handbook: Fundamentals and Applications in Contactless Smart Cards and Identification: Wiley, John & Sons Incorporated (2003)

[2] Managing RFID in Supply Chain, Int. J. Internet Protocol Technology, Vol. 2, and Nos. 3/4, 2007.

[3] Roozbeh Derakhshan, Maria E. Orlowska and Xue Li., RFID Data Management: Challenges and Opportunities 2007, IEEE International Conference on RFID Gaylord Texan Resort, Grapevine,

[4] Hector Gonzalez Jiawei Han Xiaolei Li Diego Klabjan Warehousing and Analyzing Massive RFID Data Sets, University of Illinois at Urbana-Champaign, Urbana, IL 61801, USA.\

[5] Yang Zhang, Nirvana Meratnia, Paul , A Taxonomy Framework for Unsupervised Outlier Detection Techniques for Multi-Type Data Sets, Technical Report, University of Twente, 2007.

[6] Elio Masciari and Giuseppe M. Mazzeo, Efficient Outlier Detection in RFID Trails ,ICAR-CNR Italy Source: Development and Implementation of RFID Technology, Book edited by: Cristina TURCU, ISBN 978-3-902613-54-7, pp. 554, February 2009, I-Tech, Vienna, Austria.

[7] Xiaogang, Su-Chih-Ling, Tsai, Outlier detection, Article first published online: 9 MAR 2011DOI: 10.1002/widm.19,Copyright © 2011 John Wiley & Sons, Inc.

[8] Xiaowei Xu, Martin Ester, Hans-Peter Kriegel, Jörg Sander, A Distribution-Based Clustering Algorithm for Mining in Large Spatial Databases, Proceedings of 14th International Conference on Data Engineering (ICDE'98)

[9] J. Laurikkala, M. Juhola, E. Kentala (2000), Informal identification of outliers in medical data, In: Proceedings of IDAMAP

[10] S. Guha, R. Rastogi, and K. Shim. CURE: An efficient clustering algorithm for large databases. *SIGMOD Rec.*,27(2):73–84, 1998

[11] Breunig MM, Kriegel H-P, Ng RT, Sander J (2000),] Identifying density-based local LOF: outliers, In: Proceedings of ACM SIGMOD, pp 93-104

[12] Muthukrishnan, R. Shah, J. S. Vitter (2004) , Mining deviants in time series data streams, In: Proceedings of SSDBMS.

[13] M. E. Otey, A. Ghoting, S. Parthasarathy (2 Fast distributed outlier detection in mixed-attribute data sets, 2 006), Data Mining and Knowledge Discovery, vol. 12, no. 2-3, pp 203-228.

[14] Mirco Nanni,Dino Pedreschi,Time focused clustering of trajectories of moving objects,journal of Intelligent Systems,v.27 n.3,p.267-289,novemeber 2006