RESEARCH ARTICLE OPEN ACCESS

Analysis on different Data mining Techniques and algorithms used in IOT

Shweta Bhatia[1], Sweety Patel[2]
1 Assistant Professor, Shri Ramkrishna Institute of Computer Education &Applied Sciences, Sarvajanik Education Society, Athwagate, Surat, India
2 Assistant Professor, S.V.Patel College of Management and Computer Application, sumul dairy raod, surat.

Abstract
In this paper, we discusses about five functionalities of data mining in IOT that affects the performance and that are: Data anomaly detection, Data clustering, Data classification, feature selection, time series prediction. Some important algorithm has also been reviewed here of each functionalities that show advantages and limitations as well as some new algorithm that are in research direction. **Here we had represent knowledge view of data mining in IOT.**

I. Introduction

The Internet of Things (IOT) and its related technologies can seamlessly integrate classical networks with network instruments and devices. The data in the Internet of Things can be categorized into several types: RFID data stream, address identifiers, descriptive data, positional data, environment data and sensor network data etc. ***[1].*** Today, IOT brings the great challenges for managing, analysing and mining data. In IOT systems, data quality management is a critical technology to provide high-quality and trusted data to business-level analysis, optimization and decision making. In order to improve quality of data, anomaly detection techniques are widely used to remove noises and inaccurate data. For anomaly detection, having more data means it's easier to detect an unusual event against the background of normal events ***[3].***

Data Clustering refers to grouping of data based on specific features and its value. In IOT, Data clustering is an intermediate step for identifying patterns from the collected data. It's most common process in unsupervised machine learning. Clustering methods are divided into 4 major categories such as: partitioning methods, hierarchical methods, density-based methods and grid based methods. Other clustering techniques also exist such as: fuzzy clustering, artificial neural network and generic algorithms.

The problem of Data classification is stated as: given a set of training data points along with associated label for an unlabelled test instances. Classification algorithm contain 2 phases: Training phase and Testing phase. On the basis of training data set, segmentation is done which encodes knowledge about the structure of the groups in form of target variable. Thus classification problem is referred to as supervised learning.

The feature selection is the process used to recognized pattern and allows you to identify attributes that affect quality index the most. After some initial level of experiment feature selection is preferable, identify what are attributes that affects a specific problem most and then perform data classification, time series prediction or anomaly detection more easily as it reduce the dimensionality in mining the problem. Features selection is to find a satisfactory feature subset from the candidate feature set, so that to reach an optimal classification accuracy and computing complexity control.

A time series is collection of temporal data objects, which includes characteristics such as: large data size, high dimensionality, and updating continuously. Representation, similarity measures and indexing are 3 components of time series task relies on. Time series representation reduces the dimension and it divides into 3 categories: model based representation, non-adaptive data representation and adaptive data representation. The similarity measure is carried out in proper manner such as: research directions include subsequence matching and full subsequence matching. The indexing of time series is linked with representation and similar measure tools ***[2].***

II. Anomaly detection algorithms

Anomaly detection algorithm could be either global or local. Role of different algorithms that can be used for IOT with data anomaly are nearest neighbour-based anomaly detection, clustering based anomaly detection, statistical anomaly detection, and spectral anomaly detection.

NN based algorithms assign the anomaly score of data occurrences relative to their neighborhood. Nearest-neighbor (NN) based anomaly detection are broadly used in areas such as: for finding similar

patches in images, HTM based applications for IT analytics, wireless sensor networks, etc.

NN based algorithms are: 1. k-NN Global Anomaly Score 2. Local Outlier Factor (LOF) 3. Connectivity based Outlier Factor (COF) 4. Local Outlier Probability (LoOP) 5. Influenced Outlierness (INFLO) 6. Local Correlation Integral (LOCI) ***[5].***

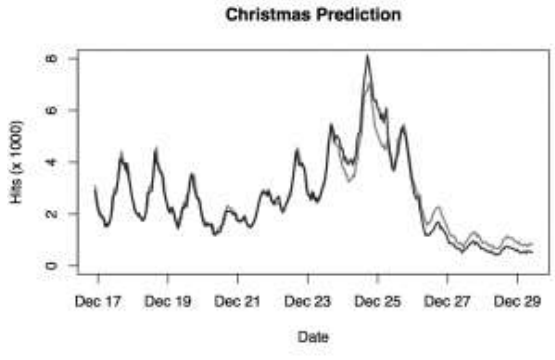
1. k-NN Global anomaly score: Score is the distance to the k-th neighbor and Score is the average distance of k neighbors. k-NN is simplest method for classification used in data mining that indirectly associate with IOT.

2. LOF: Local Outlier Factor: Most prominent AD algorithm and is able to find local anomalies. Efforts is $O(n^2)$. For example: compute the local density:

$LRD_{min}(p) = 1/ (\sum o\varepsilon N_{min} (p)$ reach_dist $_{min} (p,o) / | N_{min}(P) |)$

Statistical anomaly detection is quickly becoming a required capability in new world of the IOT. This is traceable for many special cases such as: continuous signal, discrete event timing, and user log files. Statistical anomaly detection is that you encode the patterns of what is normal as a probabilistic model. From benefit point of view, probabilistic models come with built-in measure of anomaly in terms of determining what is anomalous and this models come with a way of learning from observed data that is called a training algorithm. For instance: a rate model for web traffic anomaly detector can do a very good at predicting traffic during late December, even though it was trained only on last week of November and first week of December ***[9].***

The key property of probabilistic is the constraint that the probability of all possible things has to sum to one. On basis of this constraint, training algorithm model concentrating probability around what is normal and thus making the modelled probability of anomalies lower.



Spectral techniques attempt to find an approximate of data using combination of attributes that capture bulk of variations in data. This technique automatically perform dimensionality reduction and suitable for handling high dimensional data sets. This can be used with unsupervised setting. The disadvantage of techniques are useful only if the normal and anomalous instances are separable in lower dimensional embedding of the data and it contains high computational complexity.

III. Data Clustering, Classification and Feature Selection Algorithm

For real time IOT data stream, fast density based clustering is required which includes density based data stream clustering. This clustering grouped as density-grid based method and density based microclustering method. The main advantage of density-grid approach is its fast processing time that is independent of the number of data points and dependent only on number of cells. On other hand, density based microclustering method keep summary of clusters in microclusters and form a final clusters from them. By using advantage of both method an author in paper ***[4]***, proposed real experiment with HDC (hybrid density-based clustering for data stream)-stream algorithm. HDC-Stream only searches in potential list and if it cannot find the suitable microcluster, the data point is mapped to the grid, which keeps the outlier buffer. In future, author will focus on distributed HDC-stream density-based clustering to improve performance in IOT.

Classification widely used algorithms for IOT while mining a data on internet are: 1. C4.5 or Decision tree, 2. k-nearest neighbour algorithm, 3. support vector machine, 4. the apriori algorithm, 5. AdaBoost algorithm. These classification algorithm can be implemented on different types of data sets and on basis of performance these algorithm also used to detect the natural disasters like cloud bursting, earth quake, etc.

1. On the basis of feature values, decision tree classifies instances. For a given set S cases, C4.5 first grows an initial A. tree using divide-and-conquer algorithm as follow: tree is leaf labelled if all cases belongs to same class S or S is small. B. Otherwise, select test based on single attribute.

Some limitation this algorithm pertains are: Empty branches, insignificant branches and over fitting. Most decision tree algorithm cannot perform well with problem that require diagonal partitioning.

2. The most common role of data mining is to find frequent itemsets from transaction datasets and derive association rules. Once itemsets are obtained, it's upfront to generate association rules. **To achieve this Apriori algorithm is helpful.** This algorithm is assumes that items within transaction or itemsets are sorted in lexicographic order. The Apriori algorithm generally perform in 2steps join and prune step. It then calculates frequency only for those candidates generated by scanning the database.

As growing pressure for classification of data in urgency situation: data classification for breach response, for e-discovery, for business unity as moving towards cloud ***[8].***

There are general 3 classes of feature selection algorithms: 1. Filter methods, 2. Wrapper methods and 3. Embedded methods ***[12].***

1. To assign a scoring to each feature, filter selection method apply a statistical measure. The attribute are ranked by the score and either selected to be kept or removed from the dataset. Examples of some filter method are: chi squared test, information gain and correlation coefficient scores.
2. Wrapper methods consider the selection of set of attributes as a search problem. Score is assigned based on model accuracy where combination of features get evaluated. The search process may use different methods such as best-first search, random hill-climbing algorithm or heuristic. Example of wrapper method is recursive feature elimination algorithm.
3. When model is created, embedded methods learn which features best contribute to the accuracy of the model. Regularization method is most common that introduce additional constraints into optimization of a predictive algorithm that prejudice the model towards lower complexity. Example of this method are: LASSO, Elastic Net and Ridge Regression ***[7].***

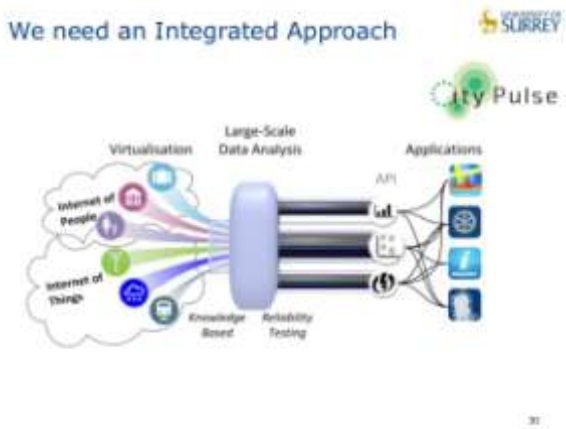
IV. Time Series Analysis and Need of IOT, challenges and conclusion.

Deep learning algorithms could apply to IOT and Smart city domains in time series analysis. The new model has been proposed with temporal patterns in deep learning namely, the **Recursive Convolutional Bayesian**

Model (RCBM), which is capable of addressing 2 tasks: **identification of multi-scale signatures and mining of compositional interactions *[10].***By building a layered structure of signature detectors, where each layer is responsible for a specific time scale, is major idea behind RCBM.

The Trendalyze Decision is also dedicated in developing and providing an analytic platform for search, visual discovery, and operational monitoring of frequently occurring patterns in time series data streams generated by IOT.

Examples of time series applications include: capacity planning, inventory replenishment, sales forecasting and future staffing levels.



To address the need for connecting large number of IOT to application infrastructure a new approach is presented called SDP developed by cloud and its public domain project available for free.

Further challenges in IOT: 1. The most practical applications are happening in Industrial IOT (IIOT) are nearly limitless such as : smarter and more efficient factories, greener energy generation, self-regulating buildings that optimize energy consumption, cities that adjust traffic patterns to respond to congestion, etc. implementation will be a challenge. 2. Security is playing vital role at all layers in IOT on the devices. There is no threat detection can mitigate effectively. Major challenges in security are: ubiquitous data collection, potential for unexpected uses of consumer data and heightened security risks. 3. IOT is not only who owns the data but who controls and receive access to that data. From a consumer perspective this will be a major challenge. 4. Shared standards and infrastructure is complex part of IOT. Structure of hardware, sensors, applications and devices that need to be able to communicate geographically and across verticals. A largest players in the market is working on developing such standards, AT&T, CISCO, IBM and Intel and GE are on the way to improve integration of physical and digital world.

V. Conclusion

This paper has focused a some algorithm on all techniques of data mining that can be applied on IOT with their advantages and disadvantages. We had also covered some future challenges and the integrated approach required to fulfil needs of IOT in present era.

REFERENCES

[1] Shen Bin , Liu Yuan* , Wang Xiaoyi*, Ningbo Institute of Technology, Zhejiang University Ningbo, China, College of Management, Zhejiang University Hangzhou, China on “Research on Data

Mining Models for the Internet of Things" in IEEE 2010.

[2] Joshua Cooper and Anne James on "Challenges for Database Management in the Internet of Things" in IETE TECHNICAL REVIEW, Researchgate.net SEPTEMBER 2009.

[3] Feng Chen , Pan Deng, Jiafu Wan, Daqiang Zhang,Athanasios V. Vasilakos and Xiaohui Rong on "Data Mining for the Internet of Things: Literature Review and Challenges".

[4] Amineh Amini, Hadi Saboohi, Teh Ying Wah, and Tutut Herawan on "A Fast Density-Based Clustering Algorithm for Real-Time Internet of Things Stream" by Hindawi in 2014.

[5] Mennatallah Amer and Markus Goldstein on "Nearest-Neighbor and Clustering based Anomaly Detection Algorithms for RapidMiner" in Researchgate.net 2012

[6] Enzo Busseti, Ian Osband, Scott Wong on "Deep Learning for Time Series Modeling CS 229 Final Project Report" in 2012.

[7] Pawan Gupta, Susheel Jain, Anurag Jain on "A Review Of Fast Clustering-Based Feature Subset Selection Algorithm" in nov-2014.

[8] Jay Cline on "Growing pressure for data classification" in 2007.

[9] Numenta on "The Science of Anomaly Detection" white paper 2015.

[10] Huan-Kai Peng*, Radu Marculescu on "Multi-Scale Compositionality: Identifying the Compositional Structures of Social Dynamics Using Deep Learning" in April 2015.

[11] Rosaria Silipo on "Data mining and Predictive Analysis" in October 2014.

[12] Jason Brownlee on "An introduction to feature selection" in 2014.