Indian J.Sci.Res. 17(2): 438-441, 2018 ISSN: 0976-2876 (Print)
ISSN: 2250-0138(Online)

ACCURATE & EFFECTIVE IMPLEMENTATION OF IOT (INTERNET OF THINGS) USING DATA MINING TECHNIQUES

J. Ravikiran
Department of Computers Science and Engineering, St.Mary's Group of Institutions, Hyderabad, Telangana

Abstract- Accurate & Effective Implementation of IOT (Internet of Things) using Data Mining Techniques is about studying the various data mining models and identify which model gives accurate results for (IoT). As an important development, the Internet of Things attracts many attentions by academic circles and industry world. IoT data has many characteristics, such asmass time-related, distributed storage and position-related data, and limited resources of nodes etc. These makes the problem of data mining in IoT become a challenge task. Implementation of Grid-based data mining systems from clustering models and the corresponding algorithms are the possibilities for the research areas.For filtering redundant data effective methods should be developed. Although there are several contributions towards data mining from IoT, they mainly focus on the rudiments of IoT, eg. RFID,sensor network etc. As a completely new paradigm of Internet, IoT is still lack of models and theories for directing its data mining.

Key words: Internet of Things, data mining models.

I. Introduction

IoT deals with connection of things globally via internet. In otherwords many applications are possible with great progress on computer communication & information technology.IoT helps in integrating communication & new computing technologies. Many researchers working in different fields like government departments, institutes & academics have shown great interest in modifying the internet by designing various systems like intelligent transportation, smart homes health care etc,. In the initial stages of work we need to review all the information about IoT device. We will extract the data stored in that device. Our next step is design the best suitable data mining algorithms which helps in connecting datamining with IoT. We may face some challenges & open issues while connecting data mining with IoT. In this paper we try to give a better solution for mining by using different data mining models.

II. Issues in Data mining of IoT:

1. A programming needs to be developed in such a way that every possible algorithm can be applied to it.

2. Framework must be designed to support security, privacy, data sharing, and data size growth etc.,

3. IoT gives high throughput, low consumption but mining algorithm is designed for low power consumption & with small size.

4. There is a possibility of creating redundant data gathering from various sources. For better system performance user needs to filter the redundant data.

5. It is very difficult to construct an architecture which connects a large number of things to the internet. Because of dependency removing of any of them may generate an error.

Parallel programming needs to be designed in such a way that every algorithm can be applied toit

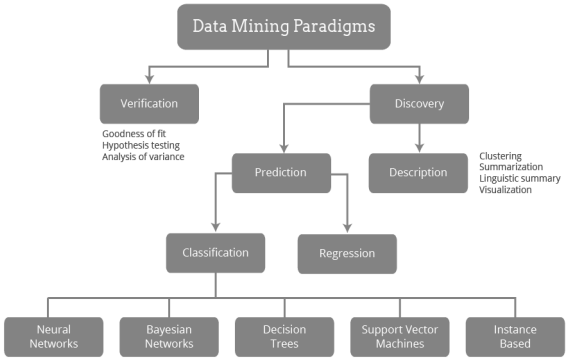
Parallel programming needs to be designed in such a way that every algorithm can be applied toit

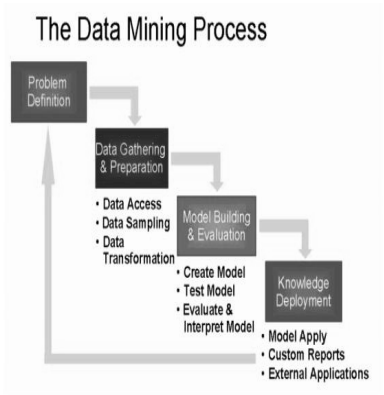
Parallel programming needs to be designed in such a way that every algorithm can be applied toit

III. Aims & Objectives:

The Internet of Things will produce large volumes of data. Let us take a University/Educational institution in anEducational domain, which adopts RFID technology for the Attendance of both Staff & Students in a regular basis, as an example. The format of raw RFID data is of the form: EPC, location, time. EPC represents the unique identifier read by an RFID reader; location is the place where the reader is positioned; and time is the time when the reading took place. It needs about 18 bytes to save a raw RFID record. In a university or any educational institution, there are about thousands of RFID tags for both students & staff. So for a RFID data stream of aninstitution, if the institution has readers that scan the items every second, about 12.6 GB RFID data will be produced per second, and the data will reach 544TB per day. Thus, it is necessary to develop effective methods for managing, analyzing and mining RFID data. The data in the Internet of Things can be categorized into several types: RFID data stream, address/unique identifiers, descriptive data, positional data, environment data and sensor network data etc.It brings the great challenges for managing, analyzing and mining data in the Internet of Things.works focus on managing and mining RFID stream data.

[1]Corresponding Author





The data mining process involves a series of steps to define a business problem, gather and prepare the data, build and evaluate mining models, and apply the models and disseminate the new information.

IV. Data mining models:

Bayesian approaches are a fundamentally important DM technique. Given the probability distribution, Bayes classifier can provably achieve the optimal result. Bayesian method is based on the probability theory. Bayes Rule is applied here to calculate the posterior from the prior and the likelihood, because the later two is generally easier to be calculated from a probability model.

One limitation that the Bayesian approaches can not cross is the need of the probability estimation from the training dataset. It is noticeable that in some situations, such as the decision is clearly based on certain criteria, or the dataset has high degree of randomality, the Bayesian approaches will not be a good choice.

The Naïve Bayes Classifier technique is particularly suited when the dimensionality of the inputs is high. Despite its simplicity, Naive Bayes can often outperform more sophisticated classification methods.

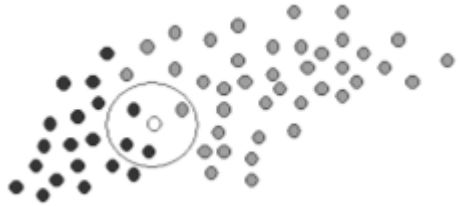
the objects can be classified as either GREEN or RED. Our task is to classify new cases as they arrive (i.e., decide to which class label they belong, based on the currently exiting objects).

We can then calculate the priors (i.e. the probability of the object among all objects) based on the previous experience. Thus:

$$Prior\ probability\ for\ GREEN \propto \frac{Number\ of\ GREEN\ objects}{Total\ number\ of\ objects}$$

$$Prior\ probability\ for\ RED \propto \frac{Number\ of\ RED\ objects}{Total\ number\ of\ objects}$$

Having formulated our prior probability, we are now ready to classify a new object (WHITE circle in Figure 2). Since the objects are well clustered, it is reasonable to assume that the more GREEN (or RED) objects in the vicinity of X, the more likely that the new cases belong to that particular color. To measure this likelihood, we draw a circle around X which encompasses a number (to be chosen a priori) of points irrespective of their class labels. Then we calculate the number of points in the circle belonging to each class label.



Since there is a total of 60 objects, 40 of which are GREEN and 20 RED, our prior probabilities for class membership are:

$$Prior\ probability\ for\ GREEN \propto \frac{40}{60}$$

$$Prior\ probability\ for\ RED \propto \frac{20}{60}$$

We can calculate the likelihood

$$Likelihood\ of\ X\ given\ GREEN \propto \frac{Number\ of\ GREEN\ in\ the\ vicinity\ of\ X}{Total\ number\ of\ GREEN\ cases}$$

$$Likelihood\ of\ X\ given\ RED \propto \frac{Number\ of\ RED\ in\ the\ vicinity\ of\ X}{Total\ number\ of\ RED\ cases}$$

it is clear that Likelihood of X given RED is larger than Likelihood of X given GREEN, since the circle encompasses 1 GREEN object and 3 RED ones. Thus:

$$Probability\ of\ X\ given\ GREEN \propto \frac{1}{40}$$

$$Probability\ of\ X\ given\ RED \propto \frac{3}{40}$$

Although the prior probabilities indicate that X may belong to GREEN (given that there are twice as many GREEN compared to RED) the likelihood indicates otherwise; that the class membership of X is RED (given that there are more RED objects in the vicinity of X than GREEN). In the Bayesian analysis, the final classification is produced by combining both sources of information (i.e. the prior and the likelihood) to form a posterior probability using Bayes Rule.

$$Posterior\ probability\ of\ X\ being\ GREEN \propto$$

$$Prior\ probability\ of\ GREEN \times Likelihood\ of\ X\ given\ GREEN$$

$$= \frac{4}{6} \times \frac{1}{40} = \frac{1}{60}$$

$$Posterior\ probability\ of\ X\ being\ RED \propto$$

$$Prior\ probability\ of\ RED \times Likelihood\ of\ X\ given\ RED$$

$$= \frac{2}{6} \times \frac{3}{40} = \frac{1}{40}$$

Finally, we classify X as RED since its class membership achieves the largest posterior probability.

Clustering is the process of making a group of abstract objects into classes of similar objects. Cluster is a group of objects that belongs to the same class. In other words, similar objects are grouped in one cluster and dissimilar objects are grouped in another cluster. As a data mining function, cluster analysis serves as a tool to gain insight into the distribution of data to observe characteristics of each cluster.

Clustering methods can be classified into the following categories

- Partitioning Method
- Hierarchical Method
- Constraint-based Method
- Density-based Method
- Grid-Based Method
- Model-Based Method

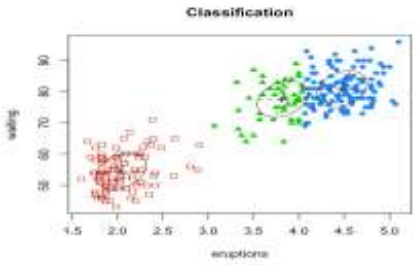
Partitioning Method

Suppose we are given a database of 'n' objects and the partitioning method constructs 'k' partition of data. Each partition will represent a cluster and $k \leq n$. It means that it will classify the data into k groups, which satisfy the following requirements −

- Each group contains at least one object.
- Each object must belong to exactly one group.

Points to remember

- For a given number of partitions (say k), the partitioning method will create an initial partitioning.
- Then it uses the iterative relocation technique to improve the partitioning by moving objects from one group to other.



Hierarchical Methods

This method creates a hierarchical decomposition of the given set of data objects. We can classify hierarchical methods on the basis of how the hierarchical decomposition is formed. There are two approaches here −

- Agglomerative Approach
- Constraint based Method
- Divisive Approach

Agglomerative Approach

This approach is also known as the bottom-up approach. In this, we start with each object forming a separate group. It keeps on merging the objects or groups that are close to one another. It keep on doing so until all of the groups are merged into one or until the termination condition holds.

Constraint-based Method

In this method, the clustering is performed by the incorporation of user or application-oriented constraints. A constraint refers to the user expectation or the properties of desired clustering results. Constraints provide us with an interactive way of communication with the clustering process. Constraints can be specified by the user or the application requirement.

Divisive Approach

This approach is also known as the top-down approach. In this, we start with all of the objects in the same cluster. In the continuous iteration, a cluster is split up into smaller clusters. It is down until each object in one cluster or the termination condition holds. This method is rigid, i.e., once a merging or splitting is done, it can never be undone.

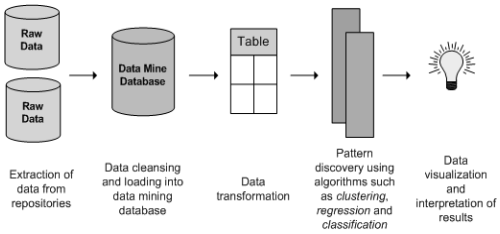
Approaches to Improve Quality of Hierarchical Clustering

Here are the two approaches that are used to improve the quality of hierarchical clustering −

- Perform careful analysis of object linkages at each hierarchical partitioning.
- Integrate hierarchical agglomeration by first using a hierarchical agglomerative algorithm to group objects into micro-clusters, and then performing macro-clustering on the micro-clusters.

Density-based Method

This method is based on the notion of density. The basic idea is to continue growing the given cluster as long as the density in the neighbourhood exceeds some threshold, i.e., for each data point within a given cluster, the radius of a given cluster has to contain at least a minimum number of points.



Grid-based Method

In this, the objects together form a grid. The object space is quantized into finite number of cells that form a grid structure.

Advantage

- The major advantage of this method is fast processing time.
- It is dependent only on the number of cells in each dimension in the quantized space.

Grid-based data mining applications are likely to use federated and existing grid technologies which hide the complexity of multiple ownership, domains, and users. However, the algorithmic approaches used at higher levels are very important for scalability and optimization of the distributed processing cost. Well-adapted algorithmic approaches are then of prime importance in the design of data mining applications and frameworks for the grid. Indeed, we are already designing and implementing the ADMIRE framework which is a data mining engine on the grid intended to provide gridbased mining techniques and dynamicity at the user level. It also provides higher knowledge map representations of the mined data, both in local nodes and globally.

Model-based methods

In this method, a model is hypothesized for each cluster to find the best fit of data for a given model. This method locates clusters by clustering the density function. It reflects spatial distribution of the data points.

This method also provides a way to automatically determine the number of clusters based on standard statistics, taking outlier or noise into account. It therefore yields robust clustering methods

Methodology :

Prerequisites:

1. Python as a programming Language, Raspberry Pi boards, Micro SD card (and an adaptor), Mini Wi-Fi adapter, Micro USB cable, Wires, Breadboard, LED, Resistors.
2. To study Data mining models the following algorithms has to be studied & Analysed:C4.5, k-means, Support vendor Machines, Apriori, EM, Page Rank, ,Naïve Bayes, Cart.
3. WEKA an Open Source Tool

V. Conclusions

In this paper, we propose grid based data mining model and its corresponding algorithms for the future work to get more effective and efficient results while mining the data from various sources.

References

[1] Cooper J, James “A. Challenges for Database Management in the Internet of Things,” IETE Tech Rev. 2009. 26:320-9.

[2] JoydeepGhosh. “A Probabilistic Framework for Mining Distributed Sensory Data under Data Sharing Constraints,” First International Workshop on Knowledge Discovery from Sensor Data. 2007.

[3] Hector Gonzalez, Jiawei Han, Xiaolei Li, Diego Klabjan. “Warehousing and Analyzing Massive RFID Data Sets,” ICDE 2006: 83.

[4] “Research on Dataminig models for theInternet of Things”,Shen Bin# , Liu Yuan* , Wang Xiaoyi* # Ningbo Institute of Technology, Zhejiang University Ningbo, China *College of Management, Zhejiang University Hangzhou, China