Session based Clustering Algorithm for data stream in Agricultural IoT

Priya Mehta[1], Jasmine Jha[2]

[1] LJIET,
Ahmedabad, India
Mpriya336@gmail.com

[2] PG Dept., LJIET,
Ahmedabad, India
jhajasmine@gmail.com

Abstract

This research explores some approaches to harnessing the IoT in agriculture. The Internet of Things (IoT) is a fast emerging system of physical sensors and connected devices, enabling an advanced information gathering, interpretation and monitoring. In this paper, we propose a new method to cluster sensor data to provide future prediction for batches data of a certain farm product and can find which method for clustering is more effective in agricultural field.

Keywords: *Clustering, Sensor Data Stream, Agricultural IoT.*

1. Introduction

The Internet of Things (IoT) is the linkage of physical objects embedded with software, electronics, network connectivity and sensors which enables these objects to collect and exchange data. The objects are allowed to be sensed and controlled remotely across existing network infrastructure with IoT. It creates opportunities for direct integration between the computer-based systems and physical world. It results in economic benefit, improved efficiency and accuracy. It is estimated that the "IoT will consist of almost 50 billion objects by 2020". Various existing technologies are used to collect useful data and it autonomously pass the data to other devices. Current examples include washer/dryers that use Wi-Fi. Traditional IoT devices were only used to transmit information, but innovation is driven to devices into existence that will convert this data into actions which will then add a totally new capability to the IoT devices and will provide better-off experience to the customers. Based on agricultural IoT, we use clustering to find similar batches of farm products that have similar status indicates the quality. Sensor data are time-sensitive and it collects different values at different times. These are considered as challenges for data mining in IoT. In this paper, because of the real time demand of agricultural IoT platform, we propose a new method to cluster the sensor data and predict the farm product quality.

2. Related Work

In agricultural IoT, different types of sensors can be equipped with different phases of product life cycle. The different phases are production, processing, transportation and marketing. A data stream is a sequence of continuously arriving data that imposes a single pass restriction where random access to the data or its brief information is not feasible [1].

There are several clustering algorithms for data stream clustering such as STREAM [2-3], CluStream [1], DenStream [2], HPStream [3], D-Stream [4], MR-Stream [5] etc. With clustering process, we can batch the data with similar statuses during its whole life cycle will have similar quality. Because of this reason, we proposed a new method to cluster the sensor data that will collect the data and with the proposed GUI it will cluster the data.

K-Means algorithm is widely used for clustering methods. In this algorithm all data points will be partitioned into K clusters according to some similarity. The advantage is its time complexity whereas limitation is that results are very sensitive to the initial value of k.

3. Proposed Method

We have used k-harmonic means algorithm for clustering in this process. K-harmonic means is proposed to overcome the weakness that occurs in the k-means in terms of the initial center point. K-means cannot work optimally at the center point of a bad initialization.

3.1 K-Means Algorithm

K-Means which is one of the most widespread clustering algorithm. K-Means was proposed by McQuenn in the year 1967. The principle of this algorithm is to partition the data into k clusters that have the similar dimension n where k is a positive integer. The process is done iteratively to obtain convergent results. Straightforward K-means algorithm is as follows:

1. Determine the value of the number of Clusters as k

2. Produce k initial centroids which is cluster centers randomly

3. Calculate the means of the current data in each cluster

4. Assign each data to the nearest centroid

5. If it is not convergent then go back to step number 3.

K-Means is an algorithm which has several advantages such as simple and easy to implement. It executes very fast in both types of clustering data small and large dimension. But the disadvantage is that results of k-means clustering is not very unique it is always changing. The reason is that k-Means algorithm is very sensitive to the early determination of the center which is random. Aside from that, k-means algorithm can only be used when attributes are of numeric type.

3.1 K-Harmonic Means Algorithm

K-Harmonic Means is a clustering algorithm that is center-based, which was presented by Zhang in 2000 and altered by Harmmerly and Eikan in 2002. K-Harmonic Means is the modification of the K-Means algorithm that can solve difficulties that arise in K- Means, and that problems are the sensitivity of the early initialization and methodologies utilized as far as its victor takes all parceling in which the relationship between the item and the closest center is extremely solid so that the enrollment of the objects won't change until they are near another center. This strong relationship will avoid the center to be able to be moved from local density data. K-Harmonic Means characterizes every group from its middle point in light of the symphonious normal computation to supplant its champ takes all methodology of K-Means portioning. Harmonic average is one of some numerical functions to calculate the average. Harmonic average is suitable to apply where the present set of records defined in relation to some units, suppose for example, to compute speed. Harmonic average of the H positive real numbers x1, x2 ...., xn can be well-defined mathematically as follows:

$$H=n/(1/x1+1/x2+\cdots+1/xn)= n/(\Sigma 1/xi) \quad xi>0 \text{ for all i}$$

In the K-Harmonic Means algorithm any k-th object is considered to be a member of all clusters to-I, with the membership function value between 0 to 1. Membership function is defined as a degree of membership of the current data to determine how probable a data become members of a cluster. The decision of the k-th objects to be the member of i-th cluster is done according to the biggest membership functions. In addition, the k-harmonic means calculation utilizes the dynamic weight capacity to each of the information when the information are a long way from any inside will be given an incredible weight and the information close to the middle will be given a little weight. This causes the k-harmonic means that is not sensitive to the early initialization. So, K-Harmonic Means can work with various initialization (both good and bad). A decent initialization is instatement which produce focuses spread in every aspect of its information. Instatement is awful when the focuses are accumulated in one region.

3. Experiment and Results

We have used air temperature and humidity, soil temperature and humidity sensors and collected farm products' sensor data in one farm with the Arduino board, transceivers and sensors as shown in fig. 1. The parameters that are considered in this research are as follows: 1.Average Soil Temperature 2.Soil Temperature Variance 3.Average Air Temperature 4.Air Temperature Variance 5.Soil Humidity 6.Air Humidity 7.Atmosphere 8.Radiation. Sensor Data was collected by microC PRO software.

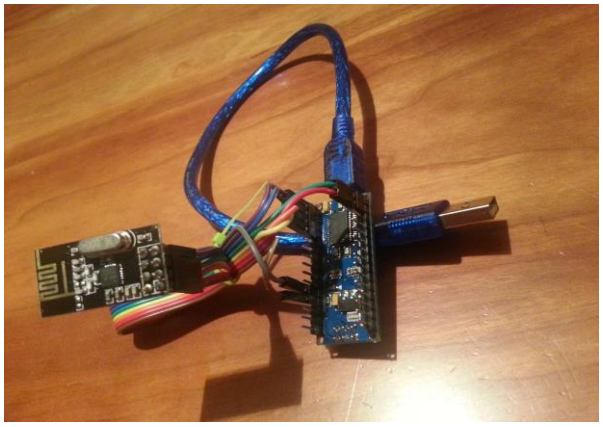

Fig. 1 Extraction of Sensor Data

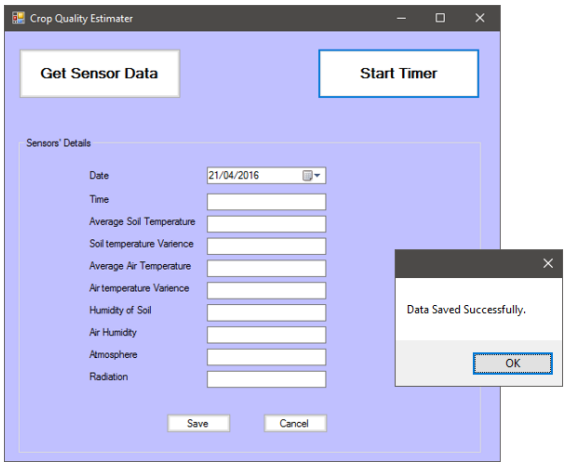

Fig. 2 Extraction of Sensor Data

The GUI for clustering is proposed in Java. Microsoft SQL is used for storing data. Figure 3 shows the GUI which is proposed for clustering.

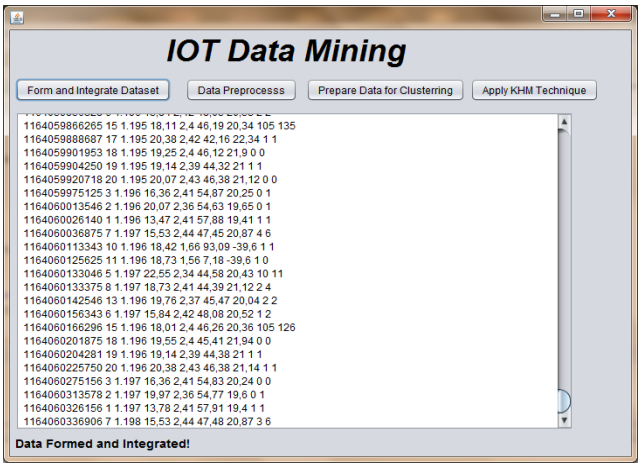

Fig. 3 GUI for clustering process

With the help of this GUI, we first preprocess the data so that noise can be removed. Then it generates .csv and .arff files that is used for clustering process.Then we apply k-harmonic means algorithm to the data set. And the results are as shown below.

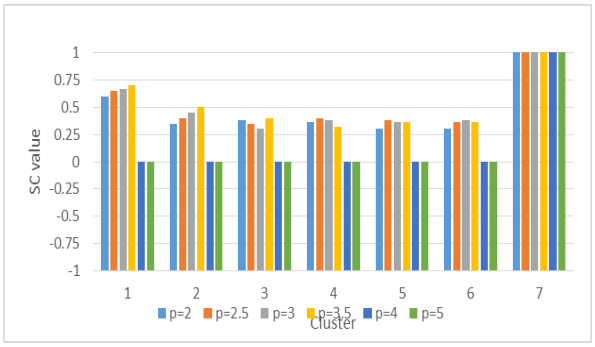

Fig.4 Comparison of k-Means and k-Harmonic means

4. Conclusions

From the figure 4, we can conclude that both the algorithms, k-mean and KHM can be applied in Agricultural sensor data. The result of k-Harmonic means with different variation of the parameter p is better than k-means. So, in the agricultural field we can predict quality of farm products better with k-Harmonic means.

References

[1] Aggarwal C, Han J, Wang J, et al. "A framework for clustering evolving data streams"Berlin, Germany: Proc of Int Conf on Very Large Data Bases (VLDB'03), 2003.

[2] Cao,F.,Ester,M.,Qian,W.,Zhou,A.: "Density-based clustering over an evolving datastream with noise." In: 2006 SIAM Conference on Data Mining, pp. 328–339, 2006

[3] Aggarwal, C.C., Han, J., Wang, J., Yu, P.S.: "A framework for projected clustering of high dimensional data streams. In: Proceedings of the Thirtieth International Conference on Very Large Databases", VLDB 2004, vol. 30, pp. 852–863. VLDB Endowment, 2004

[4] Chen Y, Tu L. "Density-based clustering for real-time stream data" Proceedings of the 13th ACM SIGKDD international conference on Knowledge discovery and data mining. ACM, 2007: 133-142.

[5] Wan L, Ng W K, Dang X H, et al. "Density-based clustering of data streams at multiple resolutions". ACM Transactions on Knowledge Discovery from Data (TKDD), 2009, 3(3): 14.

[6] Citra Lestari N1, Shaufiah2, Angelina Prima K3, k-harmonic means algorithm for clustering telecommunication customer data" October, 2010.

First Author complied Bachelor of engineering in Computer in 2014 from indus university and currently studying in LJIET and doing research on IoT and Data mining field.

Second Author currently at the position of Assistant Professor at LJIET, Ahmedabad