FINAL REPORT :- DIH SUMMER PROJECT

TITLE :- PROBABILISTIC APPROACH BASED ANOMALY DETECTION TECHNIQUES FOR REAL TIME SYSTEMS

Prepared by:-

Shikhar Gupta(15035040)

[Ceramic Engg. & 4th year]

Rahul Meena (16084012)

[Electrical Engg. & 3rd year]

Supervised By :-

Dr. Hari Prabhat Gupta

(Dept. of Computer Science)

KEY POINTS :-

- *Anomaly detection* is a identification of rare items , data and observations which is different from majority of the data.
- Due to anomalous data some kind of problem such as *bank fraud , a structural defect ,medical problems or errors in text may remain undetected*.
- *Lack of accurate, timely or relevant data from across the business* is also a major concern among companies when making major project.
- In this project we are solely focused *on Anomaly detection concerning with Sensor output data.* Our objective will be limited to detect Sensor giving anomalous output (if any) based on past precedents of it.

ANOMALY DETECTION IN SENSORS :-

- *Expanding IOT industries is rapidly using sensors*. For ex. self-driving car, auto-pilot in airplanes, self-functioning household appliances etc.
- Entire *decisions making* by these IOT applicances depend on what data their sensors are giving to them.
- *Sensor anomaly will influence decision making* & this can have *moderate to severe consequences* on the functioning of device. Ex. Speed sensor malfunctioning in self-driving car and overheating due to fault in temperature measuring devices.

OUR APPROACH :-

- For anomaly detection, we used a technique called ***Hidden markov model***. A HMM is a finite no. of states, each states linked with a probability distribution.
- Transitions among its hidden states are conducted by the set of probabilities called ***Transitions probability***.
- In a particular state a possible outcome or observation can be generated which is related to observation of probability distribution also called ***Emission probabilities***.
- The *sequence of emission probabilities* generated by an HMM gives some information about the sequence of states.

DECODING PROBLEM OF HMM :-

We are using decoding problem of HMM modelling which is as follow :

- In decoding problem we want to find state sequence with most probability of observations,

$$O = o_1, o_2, \ldots, o_T$$

and a model,

$$\lambda = (A, B, \pi).$$

- The whole state sequence with the maximum probability is found,

$$\delta_t(i) = \max_{q_1 q_2 \cdots q_{t-1}} p\{q_1, q_2, \ldots, q_{t-1}, q_t = i, o_1, o_2, \ldots, o_{t-1} | \lambda\},$$

- We observe the following recursive relationship,

$$\delta_{t+1}(j) = b_j(o_{t+1}) \left[\max_{1 \le i \le N} \delta_t(i) a_{ij}\right], \quad 1 \le i \le N, \quad 1 \le t \le T-1 \qquad (1.8)$$

CONTD.

where,

$$\delta_1(j) = \pi_j b_j(o_1), \quad 1 \leq j \leq N$$

So we find the state sequence with highest probability from,

$$\delta_T(j), \; 1 \leq j \leq N$$

Using recursion in eq. 1.8, finally the ***j**** is found.

$$j^* = \arg \max_{1 \leq j \leq N} \delta_T(j),$$

This whole algorithm can be interpreted as a search in a graph whose nodes are formed by the states of the HMM in each of the time instant ***t*** **(1<t<T).**

DATASET INFORMATION :-

- *Time series data generated from 8 MOX gas-sensors* were used for sensor anomaly detection modelling.

- The dataset has recordings of *8 gas sensors , a temperature sensor , a humidity sensor in array with respect to time*.

Attributes :-

- This data-set have 11 attributes of which sensor data is in *3-10 (total 8 chemical sensors) attributes.*

- Data is in *time series and down-sampled to 10 instances per second* for 300 seconds i.e. a total of ~2970 instances.

link for dataset :- ***https://archive.ics.uci.edu/ml/machine-learning-databases/00362/***

IMPLEMENTATION:

- Sequential steps followed in the implementation is as follows:

- First of all, *manipulation of dataset values* was done so as to make them workable i.e. from csv file to lists etc. Data belonging to each sensor was put in ‘Lists’ for both base(training) data and testing data.

- Time-Series data from each sensor was assigned *sequentially to a fixed no. of vertical blocks based on its absolute value distribution*.

- Each *instance* of the dataset was considered as *an observation*, continuous sets of *these observations were used to calculate the probability of hidden/underlying anomalous sequences* using Hidden-Markov model.

- *Deviations in no. of blocks b/w base data & test data results in the probability of anomaly of the maximum deviation sensor* for each observation to form Emission Matrix

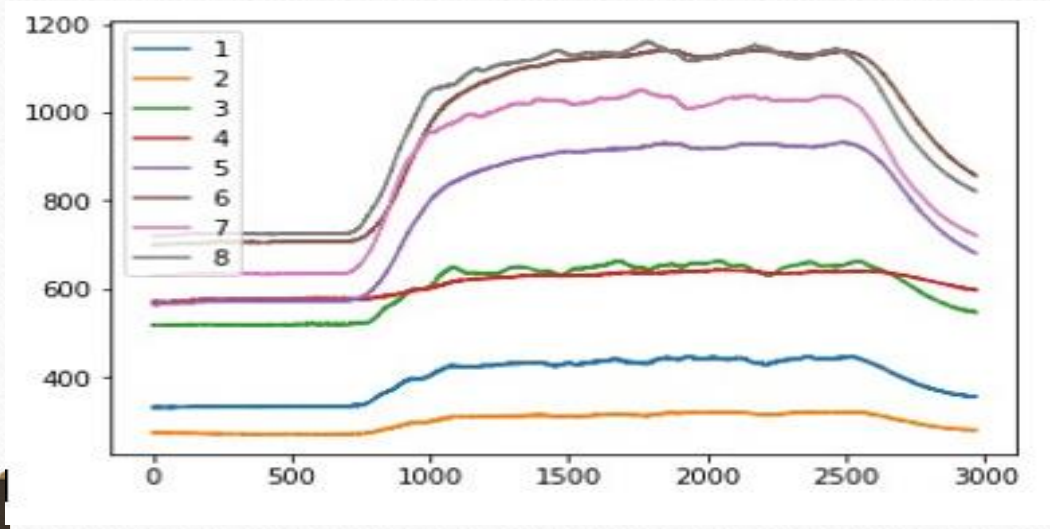
- For a given no. of *continuous such deviations, each such sensor was categorized as anomalous unless maximum deviation sensor changes*.

Code Output:-

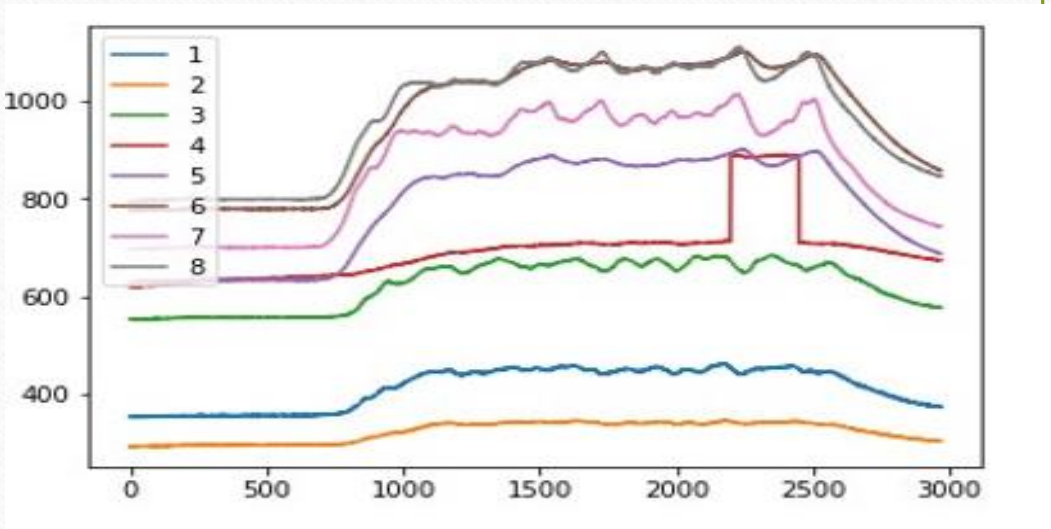
- *Visualization of sensor reading based vertical data distribution* of both training and testing data was yielded on code execution.
- *A line chart with one chart showing progression of time series of all 8 gas sensors was used*.
- Difference b/w these *vertical distributions of each sensors which were used to calculate emission probabilities* were also drawn on a line chart.
- Finally*, sequential output carrying the following items* was yielded as a result of the code execution: -

• *Anomalous Sensors (b/w 0 to 7)* caught by the Model.
• Their respective *probability of anomaly*.
• Sequential *location of the anomaly*.
• *Intensity* describing, for how long does this anomaly

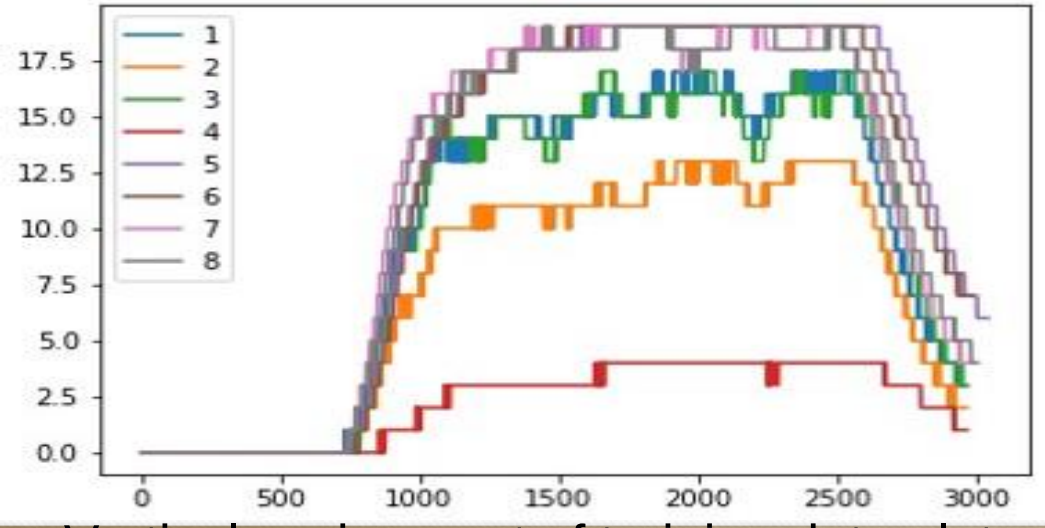
OUTPUT VISUALISATION :-



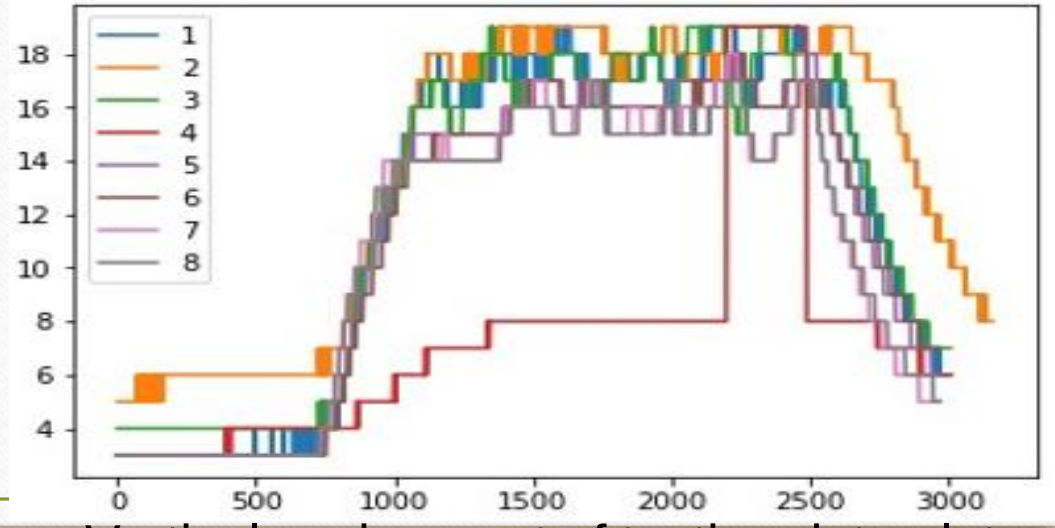
Training Data



Testing Data

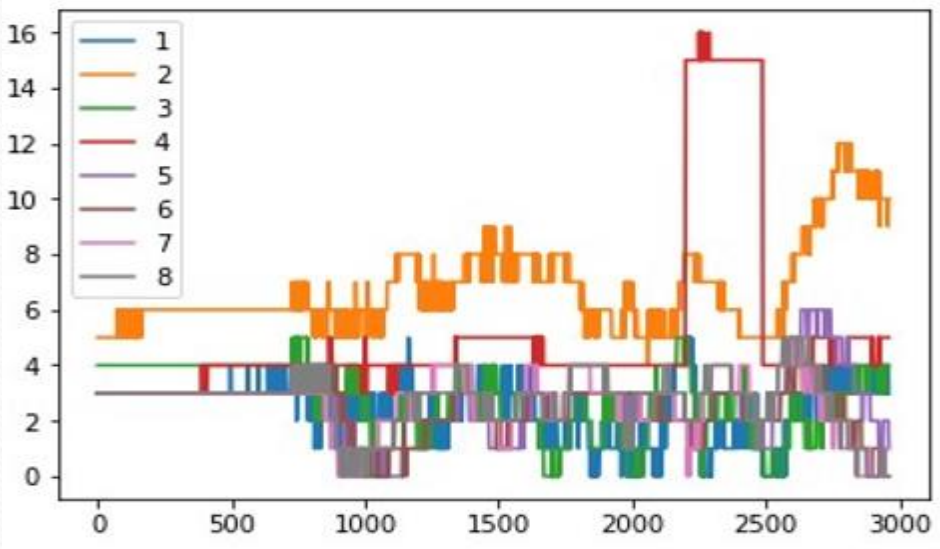


Vertical assignment of training data plot



Vertical assignment of testing data plot

CONTD.



Training and Testing data difference plot

Ex: -

Seq.	Track	Sensor no.	Intensity	Probability of Anomaly
1	0	1	219	0.0000242600
2	2198	3	28	0.1480821534
3	2487	1	47	0.0024076676

RESULT :-

Following observations were obtained out of numerous test cases that the algorithm was put through:

- *Anomaly detection becomes easier & its probability increases as % change in attributes values is increased* (say from 30% to 50%).
- *Anomaly detection is easier at the start* & medium but becomes difficult at the end of the dataset.
- *Sensors whose output value range is larger (sensor no. 5,6 & 7) are detected at a higher % increase in their values*.

THANK YOU