Unsupervised Hybrid Fuzzy Clustering Algorithm for Anomaly Detection

Akanksha Toshniwal [a], Kavi Mahesh [b], Jayashree R [c]

[a] PES University, Bangalore, KA560085, India
[b] PES University, Bangalore, KA560085, India
[c] Indian Institute of Information Technology, Dharwad, KA580009, India

Abstract

Now a days every industry no matter what domain has customer service associates. Organizations with large number of customers have lot of data and events with respect to customer services. There are some unwanted events which causes customer dissatisfaction or customer attrition. These may be the rare events which may be unfavorable to the business and remain unidentified in huge data. For example, customer calls to the customer service associate and waiting in a long waiting queue, or transferred to multiple departments etc. This manuscript provides the real time algorithm to prioritize the inbound calls (inbound calls are from customer-to-customer service associate) in order to improve customer satisfaction and engagement experience by learning their behavior and detecting anomalous inbound calls.

While working on fast increasing volumetric real-life data, it is impossible to tag each observation for modelling. This manuscript proposes an unsupervised machine learning algorithm for real time behavior learning and anomaly detection in inbound call data records. This is an unsupervised hybrid fuzzy algorithm which supports the real time untagged data.

Keywords

Anomaly Detection, Fuzzy Membership Function, Unsupervised Machine Learning, Hybrid Algorithm.

1. Introduction

In most of the business organization the biggest stakeholders are customers and it is very important to maintain the stakeholder’s interest and engagement in business. One of the biggest customer communication department is customer service department. Customer Service representatives/associates are responsible to grow, retain and improve customer experience and engagements. Call data records is one of the channels where customer directly communicate to business representatives and if something goes wrong here will result in customer dissatisfaction customer attrition. It is important to understand the call records behavior in order to ensure customer satisfaction either by prioritizing or by attend customer without delay in real time. We assume that there are rare call records which are lets say are in waiting in extremely long queue, or kept on hold for long time, call drops immediately after connection etc. The aim to detect these rare records.

With the fast growing data in every field, the possibilities of unusual activities within the data increases. These activities remain unidentified in huge data and may affect the system in long term. If the system is business, then these unusual events could cause business a huge loss. Hence, it is important to find out these rare activities.

Algorithms, Computing and Mathematics Conference, August 19 – 20, 2021, Chennai, India.
EMAIL: drkavimahesh@gmail.com (Kavi Mahesh)
ORCID: 0000-0002-4353-725X (Kavi Mahesh)

The rare or unusual activity in the dataset is called as an anomaly. The phenomena of detecting the rare event or observation from a dataset is called as anomaly detection. To perform anomaly detection, it is required to learn the patterns of the activities.

This paper proposes a machine learning hybrid unsupervised Fuzzy Clustering algorithm for anomaly detection, which detects the unusual behavior of the inbound calls based on the various factors occur in call like hold time, queue time, department to which call was made, number of transfers within a call, whether it is a weekday/weekend, whether the time of call is day/night etc.

2. Previous Work

The anomaly detection algorithms provide the rare events which occurred in groups, as point/singular event or with context (Toshniwal, 2020), (Varun Chandola, 2009). (Zadeh, 1965) was the founder and first person to work on fuzzy sets. In article by (Dey, 2009) the method to detect anomalous users and their communication patterns is proposed using fuzzy function. The communication graph is extracted with call mapping among nodes and frequency mapping as thickness to edges. (Kumar, 2018) proposes improved fuzzy membership function for feature selection for anomaly detection classification problem on KDD data, that shows the performance on low frequent classes as compared to SVM, KNN, CANN. The limitation to (Kumar, 2018) is that it is a supervised approach where labelled dataset is available but in case of the absence of the labels this algorithm fails.

3. Methodology

The input to the algorithm includes all the features in call data records. A synthetic dataset with 100k records was used to implement the algorithm. The advantage of using this methodology is that the freedom to go to granular level. For example: the hierarchy like, day of week, time of the day, department where call has been made and so on. The hierarchy is maintained for each call record and the output to the hybrid algorithm are a tagged record, where each record is tagged as anomalous or normal call. The methodology is the hybrid approach where the first step is to compute the continuous features approximate threshold values among the hierarchies and the second step is to tag the record as anomalous or normal.

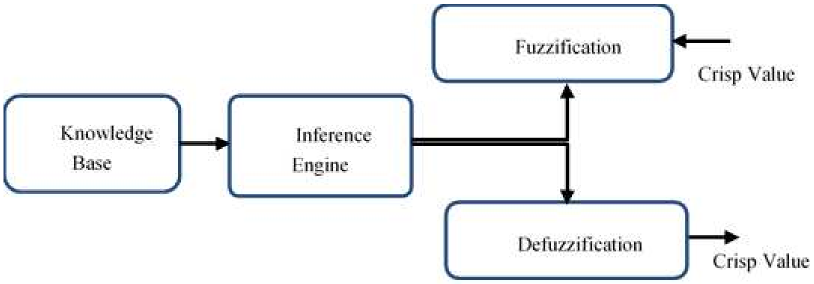


Figure 1: Fuzzy Inference System

The first step is called as Fuzzification/Defuzzification member function.

3.1. Fuzzy Systems

A fuzzy membership function characterizes fuzziness in a fuzzy set, (whether the elements in the set are discrete or continuous) in a graphical form for eventual use in the mathematical formalisms of fuzzy set theory.

The Fuzzy Inference System (FIS) is composition of fuzzification, defuzzification member functions. Figure 1 is the high-level block diagram of Fuzzy Inference System. Fuzzy Inference System is the process of formulating input/output mappings using fuzzy logic.

Fuzzy Membership Function (FMF) (Fuzzy Logic - Membership Function, 2020) is defined as the process of generating fuzzy membership values from crisp set (fuzzification) and transforming this fuzzy output of fuzzy inference system into crisp values (defuzzification). Formal definition of fuzzy membership function is given below:

Let us consider fuzzy set A,

$$A = \left(x, \mu A(x)\right) | x \in X$$

……Equation (1)

where µA(x) is called the membership function for the fuzzy set A. X is referred to as the universe of discourse. The membership function associates each element x Є X with a value in the interval$[0,1]$.

In fuzzy sets, each elements is mapped to$[0,1]$ by membership function. That is,

$$\mu A: X \in [0,1]$$

…..Equation (2)

where $[0,1]$ means real numbers between 0 and 1 (including 0,1).

There are multiple variations of fuzzy membership function. Few of them are:

a. Triangular
b. Trapezoidal
c. S-function fuzzy membership
d. Z-function fuzzy membership
e. Gaussian fuzzy membership

This is the first step of the hybrid algorithm for anomaly detection. This module consumes and produce the real-world continuous values also called as crisp values.

- Fuzzification: This step takes crisp-values or the real value as the input, and returns the real numbers between 0 and 1, see equation 1 and 2. This algorithm is implemented using the equation 3, which denotes the S-function fuzzy membership in Figure 2.

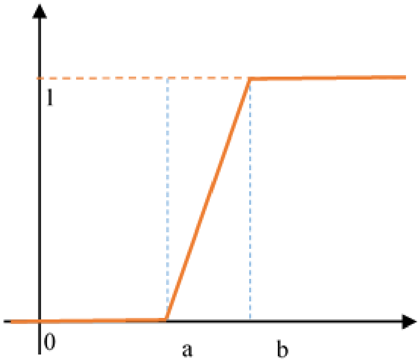


Figure 2: S-Function Fuzzy Membership

- Defuzzification: Defuzzification returns the crisp or the real values from the fuzzy membership function which goes as the input to the second step module of the ml algorithm. Figure 3 shows centroid based defuzzification (Van Leekwijck, 1999) function and equation 4 denotes the computation of centroid defuzzification function.

$$u_A(x) = \begin{cases} 0, x < a \\ \dfrac{x-a}{b-a}, a \leq x \leq b \\ 1, x > b \end{cases}$$

……Equation (3)

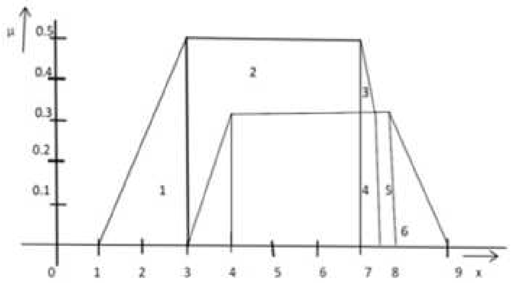


Figure 3: Defuzzification-centroid Function (Samanta)

$$x^* = \left(\sum_{i=1}^{n} x_i\, u_A(x_i)\right) \middle| \left(\sum_{i=1}^{n} u_A(x_i)\right))$$

……Equation (4)

To understand the Fuzzy Membership function, let’s have a look to an example.

Figure 4 depicts a fuzzy rule which defines vehicle speed based on the temperature and climate. Fuzzy membership function of type z, type s and triangular are used to define different temperature rules in Figure 4 (a) and fuzzy membership function of type z and triangular are used to define different weather rules in Figure 4 (b) for fuzzification and fuzzy membership function of type z and type s are used to define different speed rules in Figure 4 (c) for defuzzification. Given a temperature as 16°C and weather as sunny 25% these rules define the speed of the vehicle is 38 M/hr. An example to understand fuzzy membership function is given in Figure 4:

Fuzzy membership function:

$$Rules \rightarrow Fuzzyfication \rightarrow Fuzzy\ Values \rightarrow Deffuzification \rightarrow Crisp\ Values$$
$$Rule: Sunny(Cover)\ AND\ Temp(Hot) \rightarrow Fast(Speed)$$

Fuzzification:

$$Input: Crisp\ Value(Temperature = 16℃), Output: Fuzzy\ Value(Cool = 0.3 Hot = 0.6)$$
$$Input: Crisp\ Value(Sunny = 25\%), Output: Fuzzy\ Value(Sunny = 0.15 Cloudy = 0.35)$$

Defuzzification:

$Input$: $Fuzzy\ Values$ $(Sunny = 0.15)$ AND $(Hot = 0.6)$, $Output$: $Crisp\ Values$ $(Speed = 38\ M/hr)$

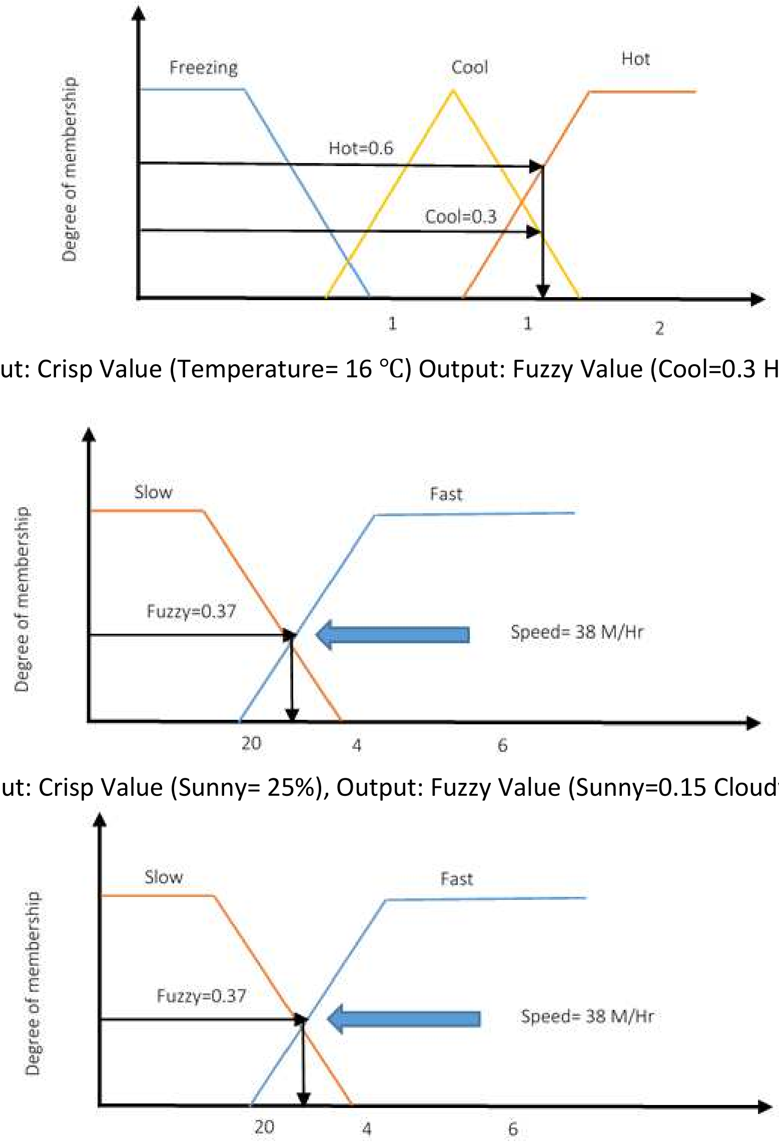


Figure 4: a) Input: Crisp Value (Temperature= 16 °C) Output: Fuzzy Value (Cool=0.3 Hot=0.6)

Figure 4: b) Input: Crisp Value (Sunny= 25%), Output: Fuzzy Value (Sunny=0.15 Cloudy=0.35)

Figure 4: c) Input: Fuzzy Values (Sunny=0.15) AND (Hot=0.6), Output: Crisp Values (Speed=38 M/hr)

Figure 4: Fuzzy Membership Function Example

3.2. Cluster Formation

Cluster formation is the second step of the hybrid machine learning algorithm. This module consumes the output from the fuzzy membership function and creates a hyper spherical cluster of radius ‘r’. This parameter is a hyperparameter in the algorithm, which is assigned as any real value. The evaluation of the instances is done based on the location of the instance in multi-dimensional space. In other words, the tagging is performed by analysing the location of the instances in multi-dimensional space. There are following steps to identify an anomalous instance which is also shown in Figure 5:

1. Any new instance which lies in hyper spherical cluster or on the cluster’s circumference is tagged as anomalous instance.
2. Any instance which lies outside the hyper spherical cluster and in quadrant Q1 are anomalous instances, see figure 5.

3. Any instance which lies outside the hyper spherical cluster and in quadrant Q2 is anomalous with respect to feature vector on Y-axis.
4. Any instance which lies outside the hyper spherical cluster and in quadrant Q4 is anomalous with respect to feature vector on X-axis.
5. Any instance which lies outside the hyper spherical cluster and in Quadrant Q3 are tagged as normal instances.

Due to the graphical limitations 3-dimenstional space and tagging in 2-dimensional plane is shown instead of multidimensional space in Figure 5. Figure 5 shows the cluster with centroid (x, y, z) with some radius ‘r’, 4-quadrants. The centroid of the hyper-spherical cluster is computed by Fuzzy Membership Function.

3.3. Evaluation

This hybrid model was implemented on synthetic Call Data Records for Customer service dataset. The inference of the results was done manually because it is an unsupervised machine learning algorithm and tagged dataset is unavailable. This technique is efficient for large data while working with granular hierarchies. Granular level represents day, time and department when the call was made, call transfer, queue time, hold time in the hierarchy order. Hence suitable for industrial solutions. Since the real tags are unavailable in the dataset, the results generated by proposed hybrid fuzzy-clustering algorithm is compared with another hybrid algorithm (clustering and xg-boost). Where it has been observed that tags associated with 99,928 records out of 100k records matched and tags associated with only 72 records out of 100k records mismatched. Which means that they both generated tagged results with 0.07% of the difference in tagging.

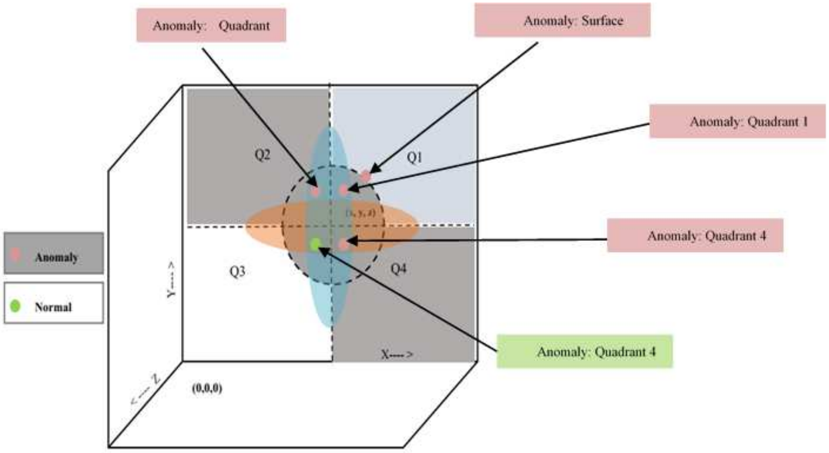
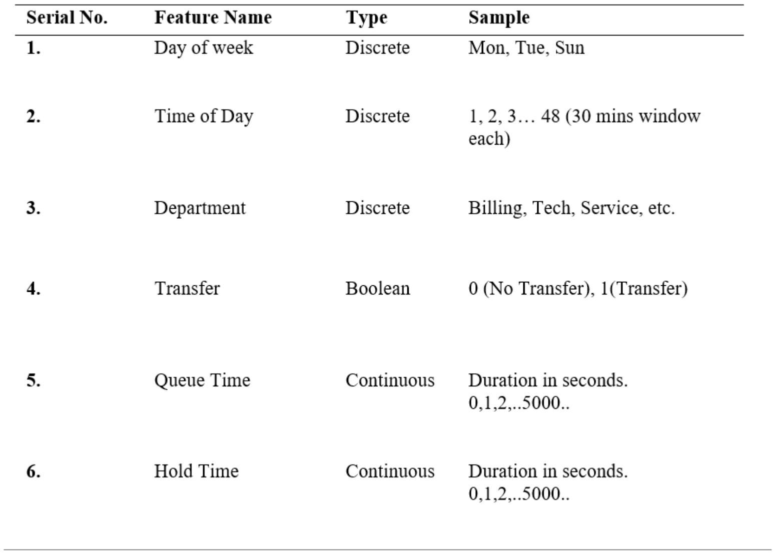


Figure 5: Tagging based on the data point position in multidimensional space

4. Dataset

A synthetic dataset of size 3 million rows was created. A random sample of 100k records was used for this experiment. There are mixed featuresin dataset. The sample dataset is provided in table 1.

Table 1: Dataset Information



Serial No.	Feature Name	Type	Sample
1.	Day of week	Discrete	Mon, Tue, Sun
2.	Time of Day	Discrete	1, 2, 3… 48 (30 mins window each)
3.	Department	Discrete	Billing, Tech, Service, etc.
4.	Transfer	Boolean	0 (No Transfer), 1(Transfer)
5.	Queue Time	Continuous	Duration in seconds. 0,1,2,..5000..
6.	Hold Time	Continuous	Duration in seconds. 0,1,2,..5000..

5. Results

This technique was implemented on the synthetic dataset for seven days of the week of size 100k size and findings are as follows:

1. On manual inference it has been observed that there are no false alarms for anomalous class.
2. The ratio of anomaly vs nominal is exceedingly small 0.001% that shows class imbalance which is quite challenging with supervised learning.
3. After comparing the proposed fuzzy-cluster based algorithm and another novel clustering-xg Boost algorithm it was observed that the rate of differences gradually decreases with increase in size of the dataset.
4. However, while implemented on granularity level, it was observed that this approach performs better with huge data.
5. Since the dataset is unlabelled, the inference is possible by comparing or by manual interpretations.
6. The real time solution is to train supervised probabilistic model with the resulting tagged data.

6. Conclusion

Novel proposed hybrid fuzzy-clustering machine learning algorithm has potential in multiple domains like anomalies in telecom, astronomy, automotive vehicles, performance measure of an application on a platform, call service centres etc. and industries like medical, business, airlines, logistics etc. This manuscript explains what anomalies are and anomalies can occur in everyday life like covid19 pandemic, attacks, wars, information security, servers, IT, vehicles, nature, weather etc.

This experiment has shown one real life application for behavior learning and anomaly detection. The proposed approach shows a potential machine learning solution to business gaps in industries. There is no existing robust solution for this gap. The fuzzy-clustering algorithm is completely unsupervised and hence is suitable for unlabelled real-life data. The advantages to this approach are it maintains the hierarchy and tags the instances as normal or anomalous among the hierarchy, proposed approach performs best on big data. There are two steps to this approach the first module uses fuzzy membership function, and the second module uses clustering.

The limitation to this approach is it does not perform well on small dataset and one should be careful while assigning the hyperparameters.

7. References

[1] Dey, L. (2009). Anomaly detection from call data records. *International Conference on Pattern Recognition and Machine Intelligence.* Berlin, Heidelberg.: Springer.

[2] *Fuzzy Logic - Membership Function.* (2020, November). Retrieved from tutorialspoint: https://www.tutorialspoint.com/fuzzy_logic/fuzzy_logic_membership_function.htm

[3] Kumar, G. R. (2018). Feature Clustering for Anomaly Detection Using Improved Fuzzy Membership Function. . *In Proceedings of the Fourth International Conference on Engineering & MIS 2018* (pp. 1-9). MIS.

[4] Samanta, D. (n.d.). *Defuzzification Methods* . Retrieved from Indian Institute of Technology Kharagpur: https://cse.iitkgp.ac.in/~dsamanta/courses/archive/sca/Archives/Chapter%205%20Defuzzification%20Methods.pdf

[5] Toshniwal, A. K. (2020). Overview of Anomaly Detection techniques in Machine Learning. *2020 Fourth International Conference on I-SMAC (IoT in Social, Mobile, Analytics and Cloud)(I-SMAC).* IEEE.

[6] Van Leekwijck, W. a. (1999). Defuzzification: criteria and classification. *Fuzzy sets and systems*, (pp. 159-178).

[7] Varun Chandola, A. B. (2009, July). Anomaly Detection: A Survey. *ACM Computing Surveys (CSUR), 41*, 1-58.

[8] Zadeh, L. A. (1965). Zadeh, Fuzzy sets. *Inform Control, 8*, 338-353.