Analysis of Clustering Uncertain Data and Uncertain Data Stream

Rao Sohail Iqbal, Ghulam Ali, Muhammad Ramzan Talib, Muhammad Awais, Shagufta Naz, Bilal Zahid, Muhammad Rehman Shahid and Muhammad Ahmad Nazir

mraosohail@gmail.com, ghulamali@gcuf.edu.pk, ramzan.talib@gcuf.edu.pk, muhammadawais@gcuf.edu.pk, shaguftanaz24612@gmail.com, bilalzahidmuavia2587@gmail.com, Mrehman0892@gmail.com,ahmad_nzpk@yahoo.com

Government College University Faisalabad, Punjab, Pakistan

Summary

The problem of handling uncertain data has been attracting the attention of researchers. This paper mainly focuses on uncertain data clustering and noise data streams. Therefore, we will provide a framework to realize the effect of uncertainty. Nowadays, a large number of tenders are present which measure the data roughly. As, sensors normally have distortion in their results cause of the imprecisions transmission in data and retrieval. Mostly these errors are identified. This information is used for minimizing process to advance results according to quality. In this paper, we compare general methods of monitor uncertainty, which have described in the different research papers.

Keywords:

Umicro clustering, Clue Stream, ELKI, SynDrift, Network Intrusion Detection, Error level, uncertain data.

1. Introduction

Recently, data sets of uncertainty have achieved significance because of its different applications in different areas [1]. Data mining is a process of pulling out hidden information from raw data and turning it out into something useful [2]. Since data collection techniques are not good and sometimes give incomplete information. For example, sensors give normally noisy data which is difficult to process for required output. There are different algorithm exist for error detection and error correction, by using these techniques we can estimate the missing data. In case of sensor stream, error occurs due to basic data collection apparatus. The results of data mining algorithm relay on the error in the data. For example, an attribute which have most error is less reliable as compare to the feature, which has small amount of error. Data types alike aforementioned will affect the results of data mining techniques [1]. In this explanation of clustering problem for uncertain data stream. Uncertainty of data mostly occurs in stream application due to austere data recording techniques generally have noisy and incomplete recording method.

This paper presents the analysis of different clustering approaches proposed by the researchers and then discusses the effective approach for the clustering of uncertain data flow. This is a big challenge because uncertainty in feature value can affect the clustering behavior [3]. Via definition cluster limit the data points touched by nature. In these days many domains like economics, healthcare, IoT, science and many other generate a large amount of data [4]. Which has very important roll to decision making of that particular fields? For this purpose, we use different algorithm to get value able result. But uncertainty occurs in data is the big challenge to obtain reliable result and gives a strong research problem. To overcome these types of problems we use different data mining process to reduce uncertainty [5].

For this purpose, we use ELKI to handle this problem, due to ELKI is an open source and developed in Java and hope that use develop algorithm in research [6]. These following steps to handle the problem of uncertain data.

- By configuring uncertain data model a no of database create from uncertain data
- ELKI is help full in clustering of uncertain data through visualization and discuss how uncertainty effect clustering result.
- Can also compare different algorithm and easily expand by user about favorite algorithm.
- Traditional clustering also implemented through sample data by an approach “Zufle”.

Over the years, we have found that a large amount of data has been developed by many applications due to a significant data increase. This paper attempts to solve the problem of uncertainty encountered in the distribution of probability between different uncertain tuples. Different methods are used to deal with this problem.

There is a problem of uncertainty in managing information as a whole. Data flow uncertain Data mining is causing great concern. This is because some applications generate a lot of data. It is not possible to interpret each tuple in memory. Therefore, we usually use algorithms to get results that guarantee the quality of the service [7]. Clustering is a more important task in data mining, some of which have different values and new tuples are adjusted according to the value accordingly.

In traditional databases, for each tuple with a single deterministic value without uncertainty, it uses the same tuple of how to handle them, the uncertain situation, to satisfy the conditions so completely I cannot do it [8]. Each opportunity represents its probability or occurrence. If two tuples A, and B are different events, for example A

Manuscript received October 5, 2024
Manuscript revised October 20, 2024
https://doi.org/10.22937/IJCSNS.2024.24.10.17

has six instances, B has two chances, since A has many uncertainties, TTS are also important. It tries to solve the problem of tracking the high quality or low uncertainty cluster specific data stream.

2. Related Work

In the work [1] the author propose the UMicro, which is called Uncertain MICRO clustering algorithm. This is used for clustering the uncertain data stream. The author make the supposition that the error of distinct items are existing. This is more accurate supposition for data mining techniques. With the nature of data mining application the description of this error can differ.The UMicro algorithm indicates and possesses a number of micro clusters. The cluster was built around gravity centers. An input, micro, indicates that the number of micro clacks is created. The algorithm moves quickly from an empty cluster and then creates a cluster. After that, new points will be added. An upcoming data point is added to the nearby cluster [1]. The latest clustering is calculated by distance between new data points and uncertain micro clusters. If data points are on the verge of uncertainty of micro clusters, they are added to micro clusters. Otherwise, a new mini-cluster will be built and contain a data point.

Clustering is most probably naïve approach is used to represents uncertain of each data clustering result is most important in filed or research [4]. It's a difficult task to handle uncertain data. Cause of that most data come from different source which most include unreliable [9]. Then interpret it much difficult task.one of most use approach that uses in many scenarios "representative clustering" which consist of 4 steps it is a fixable technique [7].

Upcoming time is the internet of things we cannot ignore this reality in [18] IoT different related devices communicate with each other, and a large amount of data generated .this data can be used fully to analyze the human behaviors and actions in society in real time [1]. The data generated through a smartphone, social networks, and smart city [5]. The Network smarter and deal an intellectual planetary to intelligence our happenings or actions and the development of the system.

Solution: we cannot perform decision-making process [2]through the personal system due to a shortage of resources.to analysis the massive amount of data through Smart Buddy. Smart Buddy is a high-level system that is used to process the large scale of data draws conclusions.it consists of three

Domain: Object domain, SIOT server domain and application domain. Smart Buddy receives the data from every data generating object like smart[3] city, body area network, social network and other and then perform data analysis the provide a direction to understand the human behavior and actions.

Parameter: there is multiple parameters in the Smart Buddy [16] system human action and behavior with IoT devices to predict the future activates and IoT devices are work like parameter like a heartbeat, temperature meter etc.

Results: part of coordinating IoT with informal communication in accomplishing human progression in light of enormous information investigation. Rather than concentrating on the information Furthermore, knowledge into information, enormous information in IoT correlations the relationship among various qualities of the information [15] and encourages the comprehension and Meaning of human conduct. Through a watchful examination of consideration and communication of natives in the rising shrewd urban areas, we have introduced the idea of Smart Buddy and shown its relevance utilizing a Hadoop biological system. Smart Buddy painstakingly handles the test of comprehension by giving input to clients that offer them the opportunity to enhance their conduct utilizing scientific categorization.

In the paper [6] ELKI framework used for design experiment and evaluating of the process. One of those method that mostly use for large data which consist of different phases. It is an open source software that mean that user can be use it according to requirement that data mining actively and continuously developed since a no of year. A no of peoples is use ELKI in related research filed such as data base and data mining. And also, where data mining process applied [6]. When we talk about technology in education system then it related to the performance is that the student performance. Lecture in old style the teacher goes in the classroom and then deliver the lecture there are many drawbacks to this method [14]. And next step the utilization of multimedia in these days has altered the method of data exchange. The certain substance is best exhibited by drawing PC illustrations which are exhibited through present day instruments like projectors [13] Support of either sounds or recordings has upset the sharing of data that have to endure for the advancement of profound learning gave the mentor examine and evacuate the misguided judgments within the same session of class. Correlation analysis performs between old lecture method and multimedia techniques. Students' key explanations behind enjoying of addressing with interactive media were increasing complete consideration, idea mapping, relating the utilization [17] of white load up with interactive media slides (all the more unmistakably clarifying) that eventually rouses students maybe for self-coordinated learning. Be that as it may [15], utilization of unadulterated interactive media without utilization of whiteboard isn't preferred by students.

Half of the instructors utilized "sight and sound just" in their addresses [11]. One-third of the instructors utilized "sight and sound alongside whiteboard", where maybe a couple of them has utilized no "media". Percent of instructors[12] who are utilizing sight and sound projectors in various modalities [18]. Demonstrate that unadulterated addressing without the utilization of sight and sound, in spite of the fact that advances self-coordinated learning, had a few impediments, for example, the misfortune of focus.

If we can calculate all the information about the tuples, the uncertainty will be zero. The probability and potential problem spaces are largely associated with their uncertainty [7]. Tuple xt is composed of two separate xT and xtb chances, and the amount of uncertainty in the declaration provided by x equals the sum of uncertainties in the declaration of xta and xtb. A Clustering Method No Remaining Data Flow Using the Micro-Clustering Framework. Aggarwal and Yu invited the features of an error-based (ECF) cluster for handling uncertain data flows.

3. Research Methodology

The research type it clearly see that it is an invitation and mostly computer science research consists on the invention, Because this provides a solution to a problem. When we come toward the research classification with different aspects which is briefly described.is it belongs to descriptive research with respect to the purpose of research and when categorized to use it lies in applied research. And doubt with time then lies in time series within longitudinal research and we collect the data in qualitative form. The factor to consider it is interesting in this sense because it is newly emerging and near future applied research that can be solved able within the specified time as well as cost. And much background knowledge available in the literature for guides and also ensure that this is not reinventing the wheel. This invitation goal of this research effective learning through e-learning.

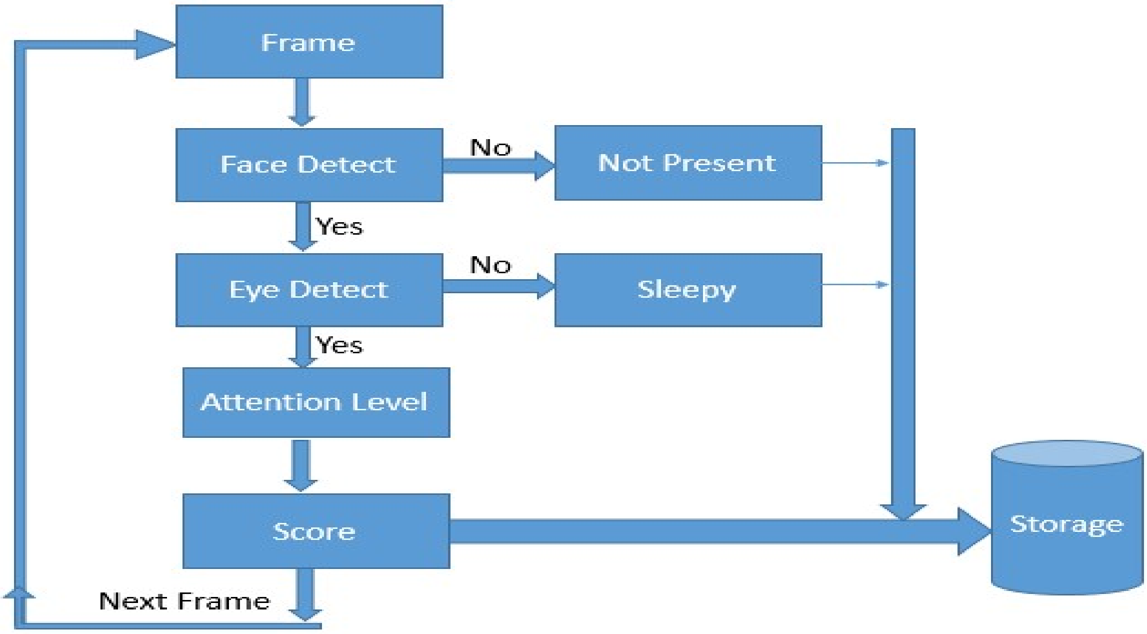


Fig. 1. Model for storage

The concept used in this research is effectiveness and attentiveness. Effectiveness has its place in the teacher performance and attentiveness be appropriate to the student. This concept measure through different dependent and independent variables, that may be dichotomous or polytomous variables. (ASM) [1] attention scoring model road model identify student basic habits behave and behavior because it cannot exist physical classroom to check the student through a video camera or webcam then else is this information to help the teacher as well as the administration to understand student learning interest through different aspects.

4. Results and Discussion

In [1], two different methods are applied to various uncertain data ranges using the Umicro clustering method and the Clue Stream clustering method. The results show that the uncertain clustering method is superior to Clue Stream clustering method. Add the noise parameter η to each data set. "Use the SynDrift dataset to use this method by storing a 600-point 20-dimensional data stream. Other data sets are network detection tests, and TCP connection records with MIT Lincoln Labs LAN network for two weeks. There are four main types of network attacks: DOS, R2L (i.e. "unauthorized access to remote machines"), U2L and PROBING [1]. These attacks are considered to be clustered. So the data has five clusters, including a simple connection type.

Fig. 2. Accuracy with Progression of Stream (Syndrift Data Set, η = 0.5)

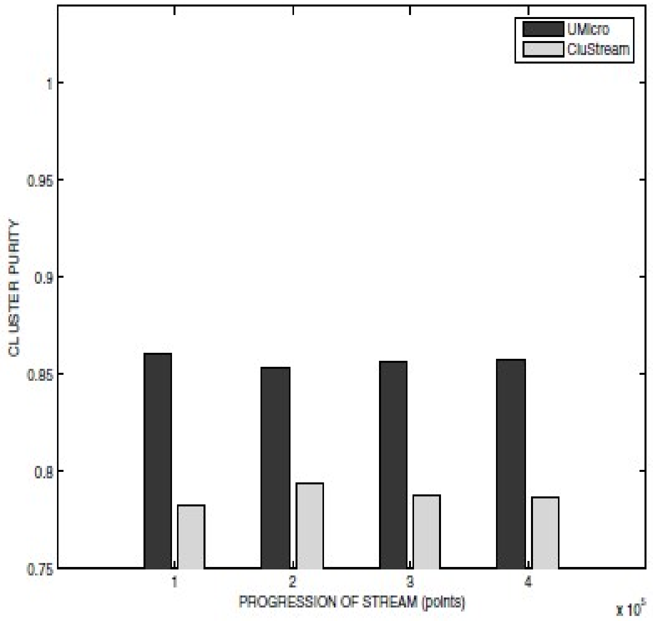
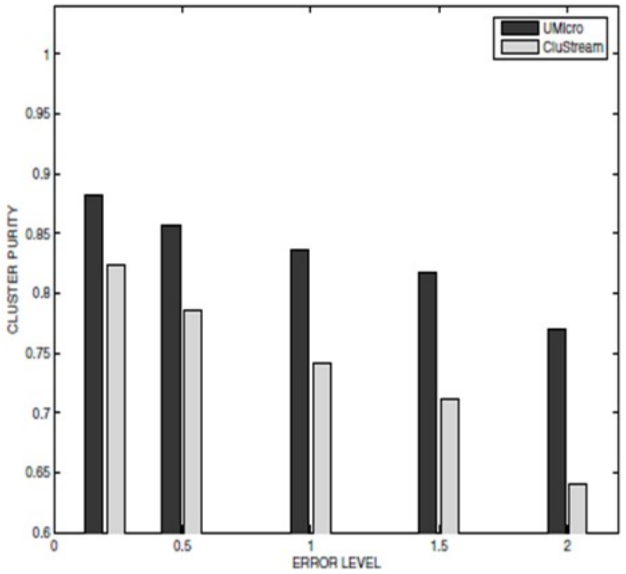
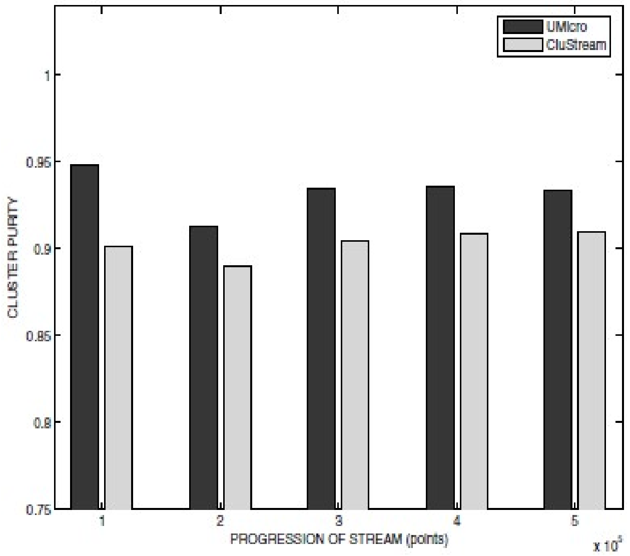


Fig. 3. Accuracy with Increase in Error Level (Syndrift Data Set)



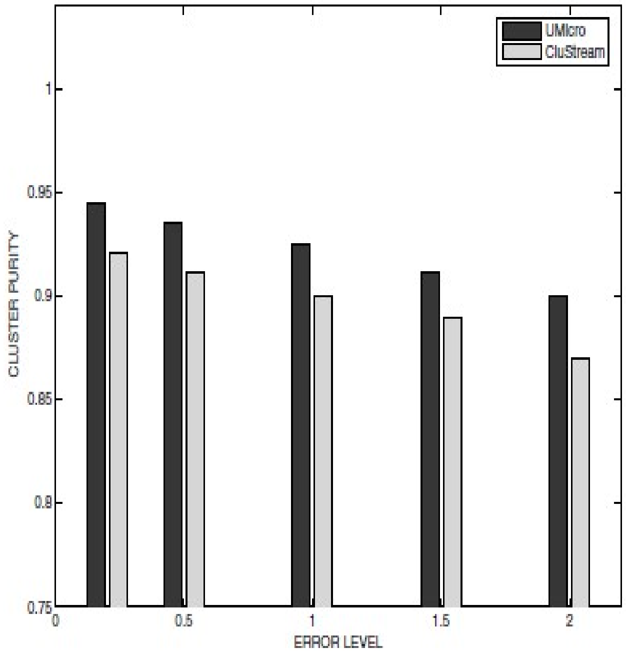
In this paper [2] calculated the clustering accuracy for constant noise and increasing level of noise on these two data sets, by using two different methods Clue Stream and Umicro method. When the noise was constant then accuracy also remains constant for these two algorithms on given data sets [9]. But accuracy decrease with respect to the increase in error level for both algorithms on *SynDrift* and Network Intrusion Detection data sets [1]. The results also shows that the Umicro clustering algorithm always perform better than the ClueStream clustering algorithm on these two data sets, wither the noise remain constant or increase in error level.

Fig. 4. Accuracy with Progression of Stream (Network Intrusion Data Set,



η = 0.5)

Fig. 5. Accuracy with Increase in Error Level (Network Intrusion Data



Set)

Here introduce our LuMicro framework which is based on PCF. With this framework, we work with online and also offline [6]. It works using iterative approach and maintains the centroid of the cluster around where cluster is generated. And assumed parameter which is predefine called cmicro.it represent the number of micro cluster generated online.it start with number of null cluster new point add by a follow a process.

The main idea is to maintain the adjustment between the position of tuple and uncertainty. For each new incoming tuple, this method expects to collect all the chances and count them as tuple values.

$$Y_t = \sum_{i=1}^{|yt|} yti * Pt(xti) \quad (1)$$

Then, it follows the measurement of uncertainty probability distribution. If there are micro clusters that can't receive tuples via remote oriented selection, they are called outliers [7]. Our algorithm inherits the same framework to handle outliers and new incoming clusters.

The Figure.2 shows that when the noise parameter remains constant, the accuracy of clustering also remains constant. The Figure.3 shows Accuracy of clustering decreases with respect to increase in noise parameter. It also shows that Umicro method perform better than CluStream method on SynDrift data. The Figure.4 shows that when the noise parameter remains constant, the accuracy of clustering also remains constant. The Figure.5 shows that Accuracy of clustering decreases with respect to increase in noise parameter. It also shows that Umicro method perform better than CluStream method on Network Intrusion Data set.

Table 1: Comparison

Author	**Journal / Conference**	**Title**	**Techniques used**	**Data Set**	**Limitations**	**Results**
Charu C. Aggarwal, Philip S. Yu	2008 IEEE 24th International Conference on Data Engineering In Cancun, Mexico	A Framework for Clustering Uncertain Data Streams	*UMicro*, the Uncertain MICRO clustering algorithm & Deterministic ClueStream Algorithm	(Syndrift Data Set, η = 0.5), (SynDrift data set with increasing error), (Network Intrusion Data Set, η = 0.5), (Network Intrusion Data Set with increasing error)	"All results, IBM T41p ThinkPad Run the 1.69 GHz operating system Windows XP Processor and 1 Tera Byte of RAM."	The result shown that in every case UMicro technique perform better than ClueStream method. The accuracy remains constant when we take noise parameter = 0.5. And decreases with increase in error level.
Erich Schubert, Alexander Koos, Tobias Emrich, Andreas Z¨ufle, Klaus Arthur Schmid, Arthur Zimek	41st International Conference on Very Large Data Bases, Kohala Coast,	A Framework for Clustering Uncertain Data	ELKI framework Naive Bayes,	file formats, such as CSV, ARFF, libSVM clustering algorithms, k-means and hierarchical clustering variations, density-based algorithms		Main target of this paper was on uncertain datamining algorithm in clustering aspects. Different parameter have different affect. If strong uncertainty in data occur, clustering are untreatable.
Chen Zhang , Ming Gao , Aoying Zhou	2009 IEEE International Conference on Data Engineering	Tracking High Quality Clusters over Uncertain Data Streams	LuMicro framework which is based on PCF	$x_t = \sum_{i=1}^{\|x_t\|} x_{t_i} * p_t(x_{t_i})$	All experiments show that Matlab v7R14 and Laptop computer with 3.0 GHz CPU 1 GB memory running on Microsoft Windows XP Professional Operating system	The result shows that in each case LuMicro method perform better than UMicro method.

Zahra Varaminy Bahnemiry, Mir Mohsen Pedram And Mitra Mirzarezaee	Applied Mathematics& Information Sciences Letters An International Journal	Clustering Uncertain Graph Data Stream	Uncertain Graph Stream Clustering UGSC Method	Real life data of karate club, where friendship between 34 members. Football club data where records of games of fall season 2000. And youtube data.	All the results were check on a Intel Core i5 and 4GB Physical memory. And algorithms were implemented in Matlab.	The results shows that the throughput depends upon the clusters size. Larger the cluster size smaller the throughput. And smaller in change in cluster if the stream are stable so less number of edges inserted and throughput is lesser.
Biao Qin, Yuni Xia, Sunil Prahakar, Yicheng Tu	2009 IEEE 25th International Conference on Data Engineering	A Rule-Based Classification Algorithm for Uncertain Data	Rule base uRule algorithm	Datasets are used diabetes, glass, iris, segment and sonar. These are existing at the UCI Repository. 9/10 of data is used for training and 1/10 for testing.	The results are checked on a PC with an Intel Pentium IV 3.2 GHz CPU and 2.0 GB main memory.	The results shows that uRule algorithm is very strong against uncertain data. The results also shows that its predication and classification accuracy reduce when we increase the uncertainty.

5. Conclusion

The main target of this paper was on uncertain data mining algorithm in clustering aspects and different parameter have different effect. If strong uncertainty in data occur clustering are untreatable. Lastly we increase complexity to apply framework. In the paper to clustering the uncertain data stream Umicro algorithm was proposed. Uncertain data streams are mostly existing in real application like sensors. That effect the clustering mechanism, due to changing value of different feature, which affect the distance measure. This paper also discusses about decay based method which is used on the data stream which change their pattern with respect to time. The results show that Umicro algorithm has more correctness than CluStream algorithm on decay base method. This accuracy grows with the growing levels of errors. Clustering data affected significantly quality of data. Introduce new matric tuple uncertainty difference tuple quality among difference PD. LuMicro algorithm has highly, processing fast and fitting with stream settings more efficiently.

6. References

[1] C. C. Aggarwal and P. S. Yu, "A Framework for Clustering Uncertain Data Stream," in *IEEE 24th International Conference on Data Engineering*, Cancun, Mexico, 2008.

[2] Z. V. Bahnemiry, M. M. Pedram and M. Mirzarezaee, "Clustering Uncertain Graph Data Stream," *AppliedMathematics& Information Sciences Letters,* pp. 85-96, 2016.

[3] S. Ding, J. Zhang, H. Jia and J. Qian, "An Adaptive Density Data Stream Clustering Algorithm," *Cognitive Computation,* vol. 8, no. 1, pp. 30-38, 2016.

[4] J. Ren, S. D. Lee, X. Chen, B. Kao, R. Cheng and D. Cheung, "Naive Bayes Classification of Uncertain Data," in *2009 Ninth IEEE International Conference on Data Mining*, Miami, FL, USA, 2009.

[5] M. Khalilian, N. Mustapha and N. Sulaiman, "Data stream clustering by divide and conquer approach based on vector model," *Journal of Big Data,* vol. 3, no. 1, p. 21, 2016.

[6] E. Schubert, A. Koos, T. Emrich, A. Zufle, K. A. Schmid and A. Zimek, "A Framework for Clustering Uncertain Data," in *41st International Conference on Very Large Data Bases*, Kohala Coast, 2015.

[7] C. Zhang , A. Zhou and M. Gao, "Tracking High Quality Clusters over Uncertain Data," in *IEEE International Conference on Data Engineering*, 2009.

[8] P. B. Volk, F. Rosenthal, M. Hahmann, D. Habich and W. Lehner, "Clustering Uncertain DataWith PossibleWorlds," in *2009 IEEE 25th International Conference on Data Engineering*, Shanghai, China, 2009.

[9] M. A. Sheela and M. C. Sunitha, "High Dimensional Data & High Speed Data Streams – A Survey," *International Journal of Advanced Research in Computer Science,* vol. 5, no. 6, p. 3, 2014.

[10] Y. Xia, B. Qin , S. Prabhakar and Y. Tu, "A Rule-Based Classification Algorithm for Uncertain Data," in *2009 IEEE 25th International Conference on Data Engineering*, Shanghai, China, 2009.

[11] M. Farhan *et al.*, "IoT-based students interaction framework using attention-scoring assessment in eLearning," *Futur. Gener. Comput. Syst.*, 2018.

[12] M. Farhan, M. Aslam, S. Jabbar, and S. Khalid, "Multimedia based qualitative assessment methodology in eLearning: student teacher engagement analysis," *Multimed. Tools Appl.*, pp. 1–15, 2016.

[13] M. Farhan, "A methodology to enrich student-teacher interaction in elearning," pp. 185–186, 2015.

[14] M. Farhan, M. Aslam, S. Jabbar, S. Khalid, and M. Kim,

"Real-time imaging-based assessment model for improving teaching performance and student experience in e-learning," *J. Real-Time Image Process.*, vol. 13, no. 3, pp. 491–504, 2017.

[15] M. M. Iqbal, M. Farhan, Y. Saleem, and M. Aslam, "Automated Web-Bot Implementation using Machine Learning Techniques in eLearning Paradigm," vol. 4, pp. 90–98, 2014.

[16] H. M. Truong, "Integrating learning styles and adaptive e-learning system: Current developments, problems and opportunities," *Comput. Human Behav.*, vol. 55, pp. 1185–1193, 2016.

[17] M. Farhan *et al.*, "A Real-Time Data Mining Approach for Interaction Analytics Assessment: IoT Based Student Interaction Framework," *Int. J. Parallel Program.*, 2017.

[18] S. Ahmad, "A New Approach to Multi Agent Based Architecture for Secure and Effective E-learning," vol. 46, no. 22, pp. 26–29, 2012.

[19] A. Paul, A. Ahmad, M. M. Rathore, and S. Jabbar, "Smartbuddy: Defining human behaviors using big data analytics in social internet of things," *IEEE Wirel. Commun.*, vol. 23, no. 5, pp. 68–74, 2016.