Industrial Engineering Journal
ISSN: 0970-2555
Volume : 52, Issue 8, No. 2, August : 2023

NEURO GENETIC WEIGHTED FUZZY C-MEANS CLUSTERING ALGORITHM FOR IOT DATA MINING: ENHANCING CLUSTERING ACCURACY IN HIGH-DIMENSIONAL DATA

Rameswara Reddy.K.V, Research Scholar, Department of Computer Science and Engineering, JS University -UP
Dr. Dhyan Chandra Yadav, Department of Computer Science and Engineering, JS University -UP

Abstract:
In the context of IoT and data mining, this paper presents a novel approach utilizing Genetic Algorithms (GAs) for clustering and optimization. The Neuro Genetic Weighted Fuzzy C-Means Clustering Algorithm (NGWFCMCA) is proposed as an effective solution for handling high-dimensional data and forming meaningful clusters. By combining a neuro-genetic algorithm with a weighted Fuzzy C-Means (FMC) algorithm, the proposed algorithm achieves improved clustering results. Furthermore, a neuro-fuzzy classifier is employed to enhance the clustering outcomes. The experimental evaluation involves the utilization of data collected from popular social media platforms such as Facebook and YouTube, as well as a mega-scale dataset from weblogs. Precision and recall metrics are employed to assess the performance of the proposed algorithm, demonstrating promising results when compared to other existing clustering algorithms. This study contributes to the exploration of the integration between IoT and data mining and introduces a novel algorithm for clustering in this domain. The findings indicate that the proposed algorithm effectively handles high-dimensional data and enhances clustering accuracy.

1. Introduction:

In this current era, we utilize IoT in diversified domains like supply management, homes, logistics, health and many more. Even we use in every personal wearables as well as in utilities. They have become a extensive and widespread in numerous organizations and infrastructures. The IoT has been novel paradigm, which links physical world, where this novel strategy could be linked to internet utilizing sensors for attaining measurements of pollution-rate, vibration, pressure, temperature, it might also define the condition of road and assist in recognizing the people in building or house utilizing RFID [1, 2, 3]. By the year 2025 [4], the amount of actuators & sensors could achieve devices of 100 billion with 3-trillion $ of revenue [5, 6, 7] that sometimes utilize protocols based on Internet known as Internet of Things. Further, this provides the probability of sensing data towards servers of cloud or it can communicate even with other kind of social-networks as stated in [8, 3].

Furthermore, in this big-data, the most well-known research topic is clustering techniques since they might discover intricate data sets structure with less knowledge. Besides, among current soft clustering strategies, to the expansion of FCM (Fuzzy c-means) algorithm, and further IFCM (Intuitionistic FCM) algorithm is been utilized extensively because of their benefits in lowering the impacts of noise and enhances the accuracy of clustering. In current algorithm IFCM, proximity degree measurement among the objects pair & determination over that aspects have been 2 significant issues that have considerable impacts on clustering outcomes.

The algorithms of hard clustering predict that crystal clear boundary presents among diversified clusters and allocate every object towards one cluster precisely. Nevertheless, in several applications in the real-time, sharp-boundary could not exist amid clusters necessarily; the object might belong towards several clusters. Because of this purpose, several algorithms of soft clustering have been researched. Moreover, the work [9] proposed a theory called rough set into clustering & projected k-means-clustering algorithm that allocates objects towards several clusters according to the lower & upper estimation of the rough-sets. Depending on FCM and FS theory that allows every data as subordinate for several clusters with diversified membership degree, which depicts the data proximity towards several centers of cluster as projected in [10].

Since conceptual bridge among rough-set& FS, the shadow-set is been implemented successfully towards clustering. For instance, k-means [12] and c-means [11] clustering algorithms is been proposed. The work [13] [14] projected 3-way clustering concept that segregates the overall region into 3 segments called boundary, positive & negative areas for depicting the 3 object states: uncertain, belong-to and not-belong-to. From the numerous popular clustering strategies, the algorithm FCM has been utilized extensively in several domains because of their maximal effectiveness and simplicity of using.

As IoT would be a most prominent novel data sources, the data science would offer a required contribution for making the implementations of IoT more effectively. The data-science has been a combination of diversified scientific domains, which utilizes machine-learning, data-mining & several other strategies for finding the patterns & novel insights from the data. Moreover, these strategies incorporate extensive range of algorithms implemented in diversified fields as in [15].

The challenging segment of smart environment based on IoT is to synthesize or choose the most suitable algorithm data-mining. Furthermore, such algorithm might generate worthable analytics, estimate further events accurately & manage services & network effectively within overall confines. In the above fig-1, it represents various applications IoT ranges from small towards large. Every block depicts IoT application, which performs definite task. Moreover, devices in environment of AAL have diversified capabilities & confines. Moreover, the raw-data generated from these devices might vary in terms of characteristics. Also, this raw-data has been converted into meaningful semantical resulting activities such as for recognizing the present state of users, needs data-mining (DM). Hence, if we process this novel data sort with conventional DM algorithms, it could not offer precise insight, Moreover, system might not establish responsive & intelligent environment. Single application of IoT highlights the devices diversity & their corresponding information.

2. Literature Review:

Several surveys over the IoT & their data has been depicted from diversified opinions. The work [16] detailed the vision & IoT characteristics from global opinion with robust & very informative discussion over 8 research domains. Moreover, it proposed architectural model for IoT for borrowing smart-phone time ie. Allowing app-store such environment for acing the authentication, uninstalling, installing & development services. The work [17] [18] depicted survey from technologies viewpoint used in IoT by possible research & implementations. The work [18] depicted generic 5-layer structure for the design of IoT system. The 5 layers from bottom towards up incorporate edge scheme, the gateway access, middle ware, internet & application, where in work [19] the research has been dedicated towards middleware of IoT. Moreover, it concentrates on communication, computing & interoperability within heterogenous environment over the applications & services. The work [20] examined the block chain strategy into a framework of IoT. Further, blockchain strategies with IoT & also depicted BCoT(Block-chain things) framework with numerous gain of the 5G connectivity.

The work [21] depicted extensive research on the mobile-crowd sourcing-research, application factors requires at the time of development, key considerations & architectures over their development. Overall above-stated surveys have been researches have been centered over architectural challenges of infrastructure of IoT, through not much concerned regarding algorithms of data-mining. Moreover, there have been more researches, which researched the data mining convergence with IoT. The work [22] presents that, considering the environment of IoT, data-mining algorithms have been proposed.

The work [23] presents technical, knowledge as well as application point of view. The work [24] [25] explored the future challenges of research because of novel big-data type, which has been heterogeneous & devices of big-data IoT. The work [24] examined the robustness of data analytics in the applications of IoT. By exploring the analytics of big data, model and strategies, they have been depicted a IoT based cloud framework. The work [24] [25] conducted research on real-time IoT streams of big-data & offered a in-depth algorithms of deep-learning & frameworks, which foster

better learning & analytics. The work [26] presents a prominent research attempts, which leveraged deep-learning & techniques assisted by cloud and fog in the environment of IoT application. Even though, the above researchers on data-mining & IoT have been robust enough & offer in-depth data mining utilization in the IoT, as they discussed the part of applications of IoT briefly.

They projected a novel hyperlink estimation approach that considers temporal-facts & utilized tensor & matrix factorizations for robust analysis. The researchers concentrated significantly on estimating the humans behavior and the utilization of temporal-records on social-networks. Nevertheless, their approach has been maximal suitable for hyperlink mining depending on features statistically. Further, projected link estimation version for intricate networks by the approach of implementing aggregation of supervised rank. Their model has been dependent on aggregation of supervised rank & has been motivated through notion, where every characteristics might provide some definite facts that might be accumulated in making maximal estimation of affiliation amid not connected entities in network.

Besides proposed a new method to social technological know-how modeling wherein the contributors themselves are prompted to discover the correlates of some human conduct outcome, consisting of domestic owner electricity usage. They proven and proved that their system allows non domain experts to collectively formulate a few of the acknowledged predictors of a behavioral outcome and that this gadget is unbiased of the outcome of hobby. They investigated the behavior of round 10,000 frequent customers of Location Based Social Networks (LBSNs) making use of their complete movement styles. They analyzed the metadata associated with the whereabouts of the users, with emphasis at the form of places and their evolution over the years. Moreover, they uncovered the patterns throughout special temporal scales for venue class utilization.

They proposed a generalized markov graph version for representing social networks and evaluated its application in social community synthesis and category. Moreover, their version depicts the diploma dissemination, the distribution of clustering coefficient & novel noticed characteristic, the traffic coefficient-distribution, have been basic for characterizing the social-network. Furthermore, specialized Markov graph approach particulars in diploma dissemination, the traffic coefficients & clustering distribution coefficient utilized in version have been 3 required information to characterize extensive social networks. Besides, Markov graph approach provides a novel view into the clustering-coefficient. Therefore, it is more appropriate for assessing the social-network.

3. Proposed method:

Genetic Algorithms (GAs) are heuristic search techniques that work based on the principles of evolution and natural genetics. Heuristic search has been performed by GAs in landscapes, multimodal, huge, & intricate, and it offers near-optimal solutions utilizing fitness-function in problem of optimization. Moreover, in GAs, search space components have been depicted as strings known as chromosomes. Such strings collection has been termed to be population. Primarily, the random population has been formed in the GA for depicting the diversified points in search-space.

A fitness function is associated with each string which provides the degree of suitability of the string. Based on the principle of survival of the fittest, the least strings are selected from the population and they are sent for reproduction. Operators such as mutation & cross-over have implemented on such strings for forming novel strings generation. Procedure of mutation, selection & crossover continues for definite amount of generations or until condition of termination has been satisfied.

3.1 Proposed Clustering Algorithm

In this contribution, analysis model for the user behavior has been projected by integrating neuro-genetic algorithm with weighted FMC algorithm for forming clusters. Such clusters have been categorized for forming more interesting clusters utilizing classifier of neuro-fuzzy. Integration of classification & clustering offers effective clusters in regard to this domain.

Seven Factor Analysis

Over the past years, social-networks have been formed by utilizing 5-factor analysis.Moreover, in such approach, 5-factor aspects called friends, qualification, gender, duration & frequency have been taken

in account. The researchers utilized some of the ranges for overall aspects in data clustering. The below table 3.1 exhibits the parameters list & ranges utilized by them. Moreover, they have been utilized the qualification aspect for simple identification in data member.

Table 3.1 Seven Factors Analysis

Parameter	**Description**	**Range of Values**
Frequency	Daily Sessions	1=one 2=two, three 3=four to six 4=more than 6
Duration	of a session	1=few minutes 2=up to one hour 3=one to three 4=more than three 5=Always online
Fr**i**ends	Number of friends	1=<10, 2=10-20, 3=20-30, 4=30-50, 5=50-80, 6=80-100,7=100-200, 8=200-400, 9=400 and more
Gender	Male or female	1=M, 2=F
Qualification	Arts or Engg.	1 = Arts, 2=Engg.
Age	Group	18 – 35 = Young Above 35 = Senior
A**r**ea	Continent	4=North America, 5=South America, 6=Australia

3.2 Neuro Genetic Weighted Fuzzy C-Means Clustering Algorithm (NGWFCMCA)

A NGWFCMCA has been projected in order to solve high-dimensional issues. In contemporary FWCM (fuzzy-weighted c-means) algorithm, the weighted-means have been computed depending on overall sample-points while in case of projected NGWFCMCA, this weighted-mean has been

computed depending on centers of cluster and remaining sample-points. Furthermore, weighted mean has been computed depending on centers of cluster, where this projected algorithm has been computationally less exhaustive when compared to contemporary FWCM.

Proposed algorithm steps are in the following
Step 1: Parameters set: N size of population, the maximal amount of T iterations, the amount of C clusters and many more.
Step 2: M chromosomes have been randomly generated, the chromosome depicts initial centers of cluster set for forming population.
Step 3: According to cluster initial centers shown by every chromosome, calculate weights for performing WFCMC. Compute the fitness of chromosome in the line by clustering outcome utilizing activation-function.

Fitness = * (1/Count of Ones) + * Sensitivity + * Specificity) Sensitivity = TP / TP + FN, Specificity = TN / TN + FP
Step 4: For every cluster, in order to carry mutation, selection & crossover operator for generating a novel group generation.
Step 5: For determining whether circumstances achieve the conditions of genetic termination, when achieved then genetic operation has been with drawn and step-6 has been proceeded, or else move to 3rd step.
Step 6: Findthe novel cluster-generation fitness; suitable individual fitness has been compared in existing cluster with effective individual fitness for finding the individual by maximal fitness.
We utilize the below formula in this algorithm for fitness function calculation
Fitness = * (1/Count of Ones) + * Sensitivity + * Specificity)
The proposed NGWFCMCA is based on the neuro genetic classification algorithm.
Here, the fitness is evaluated based on the formula given in the step 3 of the algorithm. For this purpose a population of consisting of their digits 0 and 1is generated based on the face book data. Each member of the current Genetic Algorithm population represents a competing feature subset that is evaluated to provide fitness feedback to the neural network. This is achieved by invoking neural network with the specified feature subset and a set of training data. This neural network produced is then is tested with the actual data which are analyzed already. Fitness of a chromosome is evaluated based upon the sensitivity and specificity from the validation dataset and number of features present in a chromosome. Among specificity, sensitivity and number of features, number of features has less importance. Therefore, the values for specificity (= 0.2), sensitivity (= 0.4) and (= 0.4) are assigned iinitially. Now the algorithm computes sensitivity and specificity based on these initial values and later on using true positive, false negative, true negative and false positive values. Furthermore, TP & TN have been considered as amount of records that are exactly categorized in abnormal & normal classes in respective order. Identically, FN & FP have been considered as amount of records that are categorized incorrectly in abnormal & normal classes in respective order..

4. Experimental Setup

In this section, the data collected from social media for evaluation, and the baseline methods for comparison are discussed. Two benchmark data sets provided are used to examine this proposed model for collective behavior learning. Besides, primary data set has been achieved from facebook. Next, the well recognized dataset in social-media has been Youtube-dataset. Moreover, to examine the scalability, a mega-scale data crawled from Weblog is used. The precision and recall metrics have been utilized to assesses above datasets which are defined as.

Precision = [TP / (TP+FN)] * 100 (4.1)
Recall = [TP / (TP+FP)] * 100 (4.2)

Table provides the comparison of precision & recall values of projected algorithm with the contemporary algorithms. From Table , it can be seen that the proposed clustering algorithm shows better performance in terms of precision and recall values when it is compared with the Weighted C- Means Clustering Algorithm (WCMCA) and Weighted Fuzzy C-Means Clustering Algorithm (WFCMCA).

Table 4.2 Clustering algorithms performance

	WCMCA		**WFCMCA**		**NGWFCMCA**	
Datasets	**Precision**	**Recall**	**Precision**	**Recall**	**Precision**	**Re**
Facebook	96.98	96.35	98.13	97.43	98.65	97
YouTube	96.27	95.93	97.23	97.23	97.92	97
Weblog	96.34	96.20	96.73	96.54	97.05	96

The main advantage of this proposed semi- supervised clustering algorithm is that it uses genetic algorithms and fuzzy rules to provide supervision through weight modification. This helps to measure the similarity in the neuro genetic weighted fuzzy c-means clustering algorithm dynamically based on temporal reasoning tasks including prediction and learning. Therefore, it has both training and testing data, based on which, precision and recall are calculated.

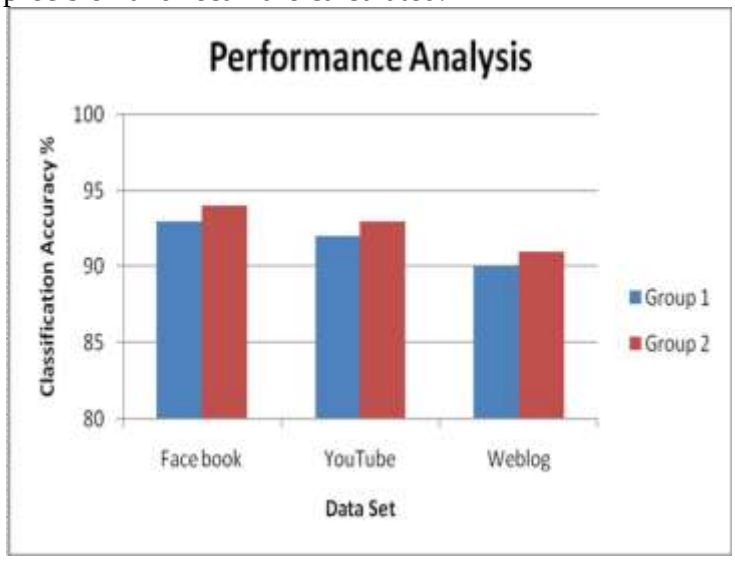

Figure 4.2 exhibis the neural classifier performance for several dataset types.

Industrial Engineering Journal
ISSN: 0970-2555
Volume : 52, Issue 8, No. 2, August : 2023

Figure 4.2: Comparison of performance analysis among group 1 &2

From Figure 4.2, the graph has been plotted among classification accuracy and diversified datasets such as youtube, weblog and facebook over the two groups. It can be observed that the classification accuracy is higher for Group2 when it is compared with Group1. This increase in accuracy for Group2is due to the behavior of the dataset.

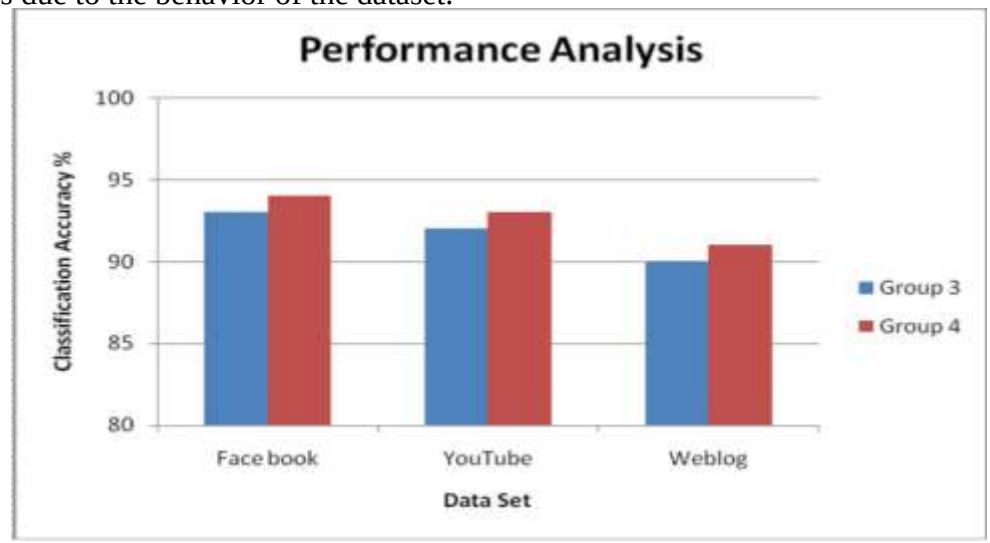

Figure 4.3: Comparison of performance analysis among group 3 & 4

From Figure 4.3, the graph has been plotted among classification accuracy and diversified datasets such as youtube, weblog and facebook over the two groups. It can be observed that Group4 classification accuracy exhibits effective than Group3. In this work four groups are considered from each of the datasets. Group 1 indicates friends, Group 2 indicates age wise group, Group 3 indicates Qualification based group and Group 4 indicates area wise group. Example: Continent.

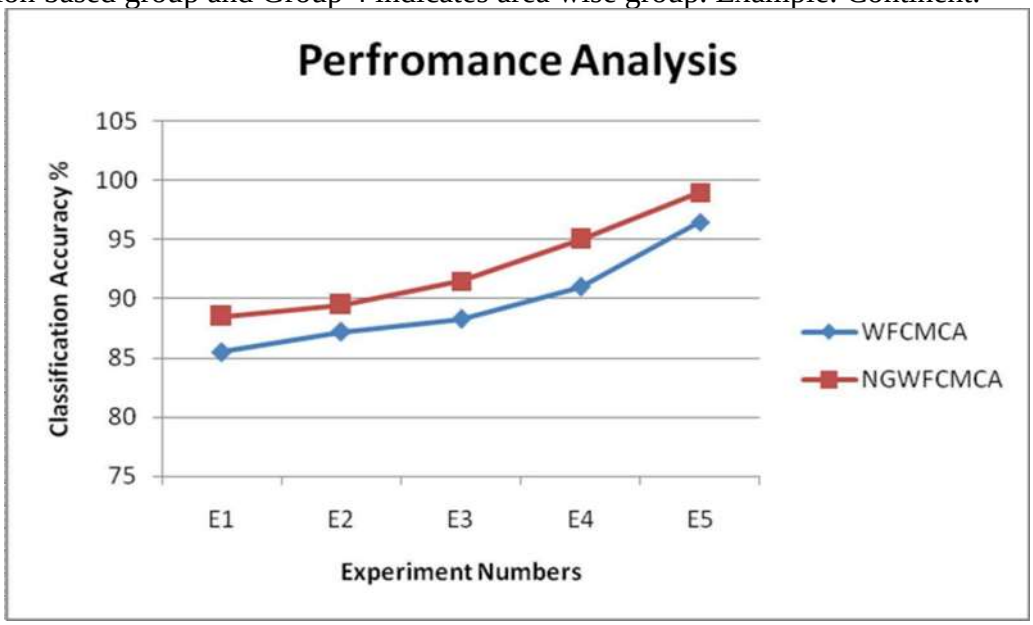

Figure 4.4 Comparing the performance on classification based on cluster

Figure 4.4,shows the overall performance analysis of neural-classifier over the several datasets with proposed clustering algorithm and the existing clustering algorithm. It is observed that the proposed model provides better classification accuracy than the existing clustering algorithm.

5. Conclusion

The proposed method introduces Genetic Algorithms (GAs) for clustering and optimization in the context of IoT and data mining. The Neuro Genetic Weighted Fuzzy C-Means Clustering Algorithm (NGWFCMCA) is presented as an effective approach for handling high-dimensional data and forming meaningful clusters. The algorithm combines a neuro-genetic algorithm with a weighted Fuzzy C-Means (FMC) algorithm, and further enhances clustering results using a neuro-fuzzy classifier. The experimental setup involves using data from social media platforms like Facebook and YouTube, as well as a mega-scale dataset from weblogs. Precision and recall metrics are used to evaluate the performance of the proposed algorithm, showing promising results compared to other clustering algorithms.

Overall, the study contributes to the exploration of IoT and data mining integration and proposes a novel algorithm for clustering in this domain. The results suggest that the proposed algorithm can effectively handle high-dimensional data and improve clustering accuracy.

References

1. M. Noura, M Atiquzaman, M Gaedke Interoperability in internet of things: Taxonomies and open challenges.Mobile Networks and Applications, 2019.
2. E Al Nuaimi, H Al Neyadi, N Mohamed Applications of big data to smart cities. Journal of Internet,2015.
3. J Gubbi, R Buyya, S Marusic, M Palaniswami Internet of Things (IoT): A vision, architectural elements, and future directions. Future Generation Computer Systems, 2013.
4. Mohsen Marjani, Fariza Nasaruddin, Abdullah Gani, Ahmad Karim,Ibrahum Abaker,Targio Hashem,Aisha Siddiqa,Ibrar Yaqoob Big IoT Data Analytics : Architecture, Opportunities, and Open Research Challenges, 2017.
5. Jayavardhana Gubbi, Rajkumar Buyya, Slaven Marusic,Maritmuthu Palansiwami. Internet of things (IOT) : A Vision, Architectural Elements, and Future Directions. Future Generation Computer Systems, 2013.
6. Summia Taj, Uniza Asad, Moeen Azhar,Sumaira Kausar. Interoperability in IOT based smart home : A review. Review of
7. Computer Engineering StudiesVol.5, No.3, September, 2018, pp. 50-55.
8. Xu, Q., Aung, K. M. M., Zhu, Y., & Yong, K. L. A Blockchain-Based Storage System for Data Analytics in the Internet of Things. Studies in Computational Intelligence, (Springer) 119–138, 2017.
9. Luigi Atzori, Antonio Iera,Giacomo Morabito. Internet of things a survey. Computer Networks Volume 54, Issue 15, 28 October 2010, Pages 2787-2805.
10. Lingras P, West C (2004) Interval set clustering of web users with rough k-means. J Intell Inf Syst Integr Artif Intell Database Technol 23(1):5–16
11. Bai C, Zhang R, Qian L, Liu L, Wu Y (2018) An ordered clustering algorithm based on fuzzy c-means and PROMETHEE. Int J Mach Learn Cybern 10(6):1423–1436
12. Mitra S, Pedrycz W, Barman B (2010) Shadowed c-means: Integrating fuzzy and rough clustering. Pattern Recogn 43(4):1282–1291
13. Zhou J, Lai Z, Miao D, Gao C, Yue X (2020) Multigranulation rough-fuzzy clustering based on shadowed sets. Inf Sci 507:553–573
14. Yu H (2017) A framework of three-way cluster analysis. In: International joint conference on rough sets (IJCRS 2017). Springer, pp 300–312
15. Yu H, Zhang C, Wang G (2016) A tree-based incremental overlapping clustering method using the three-way decision theory. Knowl Based Syst 91:189–203
16. Priyanka, E.B., Maheswari, C. and Thangavel, S., 2018. IoT based field parameters monitoring and control in press shop assembly. Internet of Things, 3, pp.1-11.

17. Stankovic, J.A., 2014. Research directions for the internet of things. IEEE Internet Things J. 1 (1), 03–09.
18. Miorandi, D., Sicari, S., De Pellegrini, F., et al., 2012. Internet of things: vision, applications and research challenges. Elsevier J. Ad Hoc Netw. 10 (7), 1497–1516.
19. Bandyopadhyay, D., Sen, J., 2011. Internet of things: applications and challenges in technology and standard. Wireless Pers. Commun. 58 (1), 49–69.
20. Razzaque, M.A., Milojevic-Jevric, M., Palade, A., et al., 2016. Middleware for internet of things: a survey. IEEE Internet Things J. 3 (1), 70–95.
21. Dai, H., Zheng, Z., Zhang, Y., 2019. Blockchain for internet of things: a survey. IEEE Int. Things J. 6 (5), 8076–8094. https://doi.org/10.1109/JIOT.2019.2920987.
22. Phuttharak, J., Loke, S.W., 2019. A review of mobile crowdsourcing architectures and challenges: toward crowd-empowered internet-of-things. IEEE Access 7, 304– 324. https://doi.org/10.1109/ACCESS.2018.2885353.
23. Tsai C.W., et al., 2014. Data Mining for Internet of Things: A Survey. IEEE Communication Surveys and Tutorials. (16)1, 451.
24. Chen, Q. et al., 2019. A survey on an emerging area: deep learning for smart city data. IEEE Trans. Emerg. Top. Comput. Intell. 3 (5), 392–410. https://doi.org/ 10.1109/TETCI.2019.2907718.
25. Marjani, M., et al., 2017. Big IoT Data Analytics: Architecture, Opportunities, and Open Research Challenges. IEEE Access (5), 5247-5261.d oi: 10.1109/ ACCESS.2017.2689040.
26. Mohammadi, M., Al-Fuqaha, A., Sorour, S., Guizani, M., 2018. Deep learning for IoT big data and streaming analytics: a survey. IEEE Commun. Surveys Tutorials 20 (4), 2923–2960. https://doi.org/10.1109/COMST.2018.2844341.