International Journal of Engineering and Advanced Technology (IJEAT)
ISSN: 2249-8958 (Online), Volume-9 Issue-2, December, 2019

Micro Clustering Methodology for Document Objects using Deep Learning Techniques

Amirkhan R. Mulla, Sachin S. Patil

Abstract: ***Large data clustering and classification is a very challenging task in data mining. Various machine learning and deep learning systems have been proposed by many researchers on a different dataset. Data volume, data size and structure of data may affect the time complexity of the system. This paper described a new document object classification approach using deep learning (DL) and proposed a recurrent neural network (RNN) for classification with a micro-clustering approach.TF-IDF and a density-based approach are used to store the best features. The plane work used supervised learning method and it extracts features set called as BK of the desired classes. once the training part completed then proceeds to figure out the particular test instances with the help of the planned classification algorithm. Recurrent Neural Network categorized the particular test object according to their weights. The system can able to work on heterogeneous data set and generate the micro-clusters according to classified results. The system also carried out experimental analysis with classical machine learning algorithms. The proposed algorithm shows higher accuracy than the existing density-based approach on different data sets.***

Keywords: Document Classification, NLP, Deep Learning, RNN, Micro Clustering.

I. INTRODUCTION

Classification is fundamental in information mining as well as machine learning. Text classification is a classification of text into some predefined categories according to its content [1]. Over the past ten years, document-based management information has become increasingly popular in the field of information systems, as there is a greater availability of documents in digital form and consequently the need to access in a flexible and efficient way. Difficulties of document classification are comprehensively analyzed and discussed in many real-life applications [10]. In the past decades, particularly with new discoveries in NLP and document mining, numerous developers involved in producing applications that advantage document classification methods. Document classification systems can follow four aspects: feature removal, dimension reductions, classification technique, and evaluations [22]. Document classification is remarkable ML task where assigns a label to the object, the main problem of multi-label clustering is repetitive clustering approach for online and offline dataset to manage this problem employ density based reclustering of current micro-clustering instances and increase the efficiency of end sub clusters [11]. Classification methods used four types of levels as follows [22]:

- Document: algorithm takes the relevant sections of a whole record.
- Paragraph: algorithm takes the related parts of an individual paragraph (a part of a document).
- Sentence: algorithm takes the suitable sections of a particular sentence (a portion of a paragraph).
- Sub-sentence: algorithm takes the suitable types of sub-expressions within a sentence (a part of a sentence).

A. Feature Removal

Text and documents both are unorganized datasets. Nevertheless, these unorganized text strings turned into well-organized features while working with analytical modeling as the role of a classifier. In initial step, the data should be refined to eliminate unwanted characters and words. Once data cleaned, conventional Feature Removal techniques can be applied. The standard methods of feature removal are Word2Vec, TF, TF-IDF, Global Vectors for Word Representation (GloVe) [22]. from above, the TF-IDF is easy to estimate the similarity between two documents and common words do not affect the final result.

B. Dimension reduction

Document or text datasets usually include multiple different words. Pre-processing methods can slow due to memory and time complexity [21]. A standard solution to this problem is merely practicing cheap algorithms. Nonetheless, in datasets, this weak algorithm not expected to perform well. To overcome the performance issue, various researchers propose dimensionality reduction to overcome the memory and time complexity for their tools [1]. It could be beneficial to use dimensionality reduction for pre-processing than making bad classifiers.

C. Classification Techniques

Picking the best classifier is the most crucial step of document and text classification. Without prior knowledge of any algorithm, we never adequately figure out the effective model for document classification.

Revised Manuscript Received on December 30, 2019.
* Correspondence Author
Amirkhan R. Mulla*, Department of Computer Engineering, Rajarambapu Institute of Technology, Rajaramnar, India Email: mulla.amir@gmail.com
Sachin S. Patil, Department of Computer Engineering, Rajarambapu Institute of Technology, Rajaramnar, India Email: sachin.patil@ritindia.edu

© The Authors. Published by Blue Eyes Intelligence Engineering and Sciences Publication (BEIESP). This is an open access article under the CC BY-NC-ND license (http://creativecommons.org/licenses/by-nc-nd/4.0/)

Recently, deep learning methods have accomplished better outcomes in comparison to prior machine learning algorithms on jobs such as NLP, face recognition, document classification [10]. These deep learning algorithms progress is depending on model complexity and non-linear relations within data.

D. Evaluation

Evaluation is the last point of the document classification. Knowing how a planned work performs and is vital to use an expansion of a document classification system. Accuracy calculation is one of the easiest ways among the various methods of evaluation [22].

The planed work concentrates on the classification of documents using Recurrent Neural Network, the system working with a 70-30% PDF dataset, respectively, for training and testing. The NLP based method used for extracting features information in training and testing. RNN allot more weights to the past data points of sequence [3]. So, this method is a dominant method for document, text, object, and subsequent data classification. In RNN, net analyses the knowledge of past neuron in an advanced scheme that provide more useful semantic analysis of structures in the dataset [5].

II. LITERATURE REVIEW

Deep learning networks require convoluted modules for multilingual sentiment polarity detection and the domain classification processed in [1] The primary purpose which highlights the capacity to build deep learning models of emotion polarity review and subject categorization in a multilingual background outwardly earlier reason. To this purpose, the researcher gained textual data of the web, penned in French, English, and Greek (low-opinion language). system used a couple of different deep neural networks, the Convolutional Neural Network and the Recurrent Neural Network. Proposed classification malware [2] using convolutional gated neural networks, the system runs malware or mischievous software, is a significant threat to the IT community. In this paper, the author proposed a convolutional gated RNN design that can classify malware into their respective families. This model has implemented a collection of malwares divided into nine different classes and offered during the Microsoft Malware Classification Challenge in 2015.Planned a system of actions for daily living classification using recurrent neural networks [3], addressing the problem of classifying an individual's daily life outside of the data based on a sensor. The system proposes to use a neural network frequently to detect continuous sensor data input and to solve problems related to long-term memory cells in the data that the activity monitors. Recurring neural network models were implemented using the TensorFlow library. LSTM RNN [4] For high-resolution range profile-based radar target classification, it is important to set a positive and timely target of the system in any military situation. The paper describes the analyze the utility of long-term memory recurrent neural networks (LSTM RNNs) based on the organization of radar targets based on high-resolution range profiles (HRRPs). Simulated radar range profile data used in this general chat lounge. This study considers three different target models classified using LSTM RNN. Proposed automatic modulation classification system using RNN [5] that implements AMC, an imperative technology in the field of cognitive radio (CR) and a non-cooperative communication system. The author proposes a new AMC system depend on the frequently occurring NN (RNN), showing sufficient ability to utilize the ephemeral flow of the obtained transmission signal. This system directly resorts to raw signals with short data length and bypasses the manual removal of signal features. System object recognition has been proposed [6] utilizing cellular symmetric RN and CNN, and they have done a lot of work from a computer perspective. paper describes, cellular symmetric recurrent network (CSRN) is used to construct a primary filter of convolutional networks (CN) and are used to classify Extreme Learning Machines (RELMs) used for regularization. In addition, CNN with DEL Belief Network (DBN), random and Gabor filters are performed to estimate the overall achievement against CN based filter based CNR with RELM. Paper [7] describe classification methods of actual data is provided using convolutional and RNN. In the past several years, artificial intelligence researchers and experts have achieved remarkable results by studying natural language processing processes, such as the convolutional and recurrent neural networks, as a result of visual, sound and new methods and deep learning. The purpose of this work is to explore, experiment and provide a new approach to the classification of non-stationary data using the neural network mentioned above. Energy consolidation system [8], NILM is the immeasurable ways to degrade electricity expenditure. Several algorithms have studied in this field. Nevertheless, classification outcomes in those algorithms are not as expected. The author proposed a new strategy for creating a taxonomy for energy dissatisfaction in the field of deep learning. The author applies a Gated Recurrent Unit (GRU) based on a recurring RNN to train the pattern utilizing the UK DALA dataset. Comprehensive Attention Recurrent Neural Network (CA-RNN) [9] that reserved, prefigure, succeed, and develop sequences in any position in the local context. Bidirectional recurrent neural network (BRNN) helpful to improve historical and upcoming information during a convolutional layer is used to taking spatial knowledge. To enhance efficiency of the new architecture, standard RNNs have also been recovered with two newly developed RNN alternatives, Long-Term Short-Memory Memory (LSTM) and Gated Recurrent Unit (GRU). Another characteristic of this planned work is self-trained outwardly any personal interference It is effortless to implement. Proposed Predicting variations in system text working with convolutional NN and clustering [10], the system analyzes text sections in some longer documents and finds articles that are differently stylized than others. The author developed two methods clustering and classification of segments using CNN and method also tested on Arabic and English long texts. The first data stream clustering algorithm that correctly registers the thickness in the zone shared by micro-clusters and applies this knowledge for re-clustering in [11].

In an experiment using 10,000 data points for clustering, imitate the experiment for each value of the ten times and report the score. Dataset Used: Chameleon dataset DS3, DS4, kdd'99. the paper described the shared density graph and analyzed the Online data stream clustering component is set to create a small number of large micro clusters. Proposed a new stream clustering algorithm called evoStream which Offered evolutionary optimization in the idle times of the online phase to incrementally develop and improve the final clusters. Dataset Used: Power supply, Sensor KDDCup'99, EvoStreams among the better for fast computation time algorithms and Deals with immortal and fast-changing data stream in [13], It refers that the clustering model extracted in species of information is also subject to evolution over time. CluStream achieves good clustering quality for several micro-ratios of 20 also observed DenStream produces better clustering quality relating to Clustream and ClusTree for every window size and throughout the entire stream execution. ACSC used to identify the clusters as large-thickness zone divided by small-thickness zone. In [14], Artificial ants sort clusters by probabilistically choosing and separating micro clusters based on the local thickness and local identity. ACSC generates good and scalable clusters it's also robust to noise and convenient to leading ant clustering and stream-clustering algorithm, and it requires fewer parameters and shorter calculation time. CluStream and DenStream in [15] Two metrics are used to estimate the accuracy (Rand Index) and efficiency (Execution Time) of the MicroGRID approach. The MicroGRID is faster up to 87% and more accurate, up to 80% clustering outputs. [16] Proposed various clustering modules 5NF, DBSTREAM, real-time clustering element that correctly caches the thickness amongst the sub-clusters through a shared density chart. Next offline step to re-cluster the micro clusters into large, very last clusters. Dataset Used: Online dataset from social media platform and Offline datasets such as a group of files or documents. Proposed DBSTREAM [17] with a median ARI Performs best in utilized several standard data sets of both real and synthetic data to recognize the fundamental gaps and strengths of the existing system. Restricted the number of micro-clusters and detailed parameter modulate enabled us to obtain a decent comparison among the various algorithms. Also observed, algorithms cannot summarize the stream adequately enough because of the limitations on the number of grid cells. The extended D-Stream algorithm increases the cluster feature but can also have the reverse effect. Introduced a research tool that includes modelling and simulating data streams [18] extensible framework used for implementing, interfacing and experimenting with algorithms for several data stream mining tasks. Dataset Used- KDD Cup'99 dataset used. Extracted 1500 data points from the Bars and Gaussians data stream generated with 5% noise. DBSTREAM [19] the first micro-cluster based on the online clustering component capturing the density between micro-cluster via shared density graph. Performance improved for the micro-cluster based system used the idea of a shared density chart that catches the density of the existing information within micro clusters during cauterization and showing how can graph used for re-clustering micro-clusters to reduce the density. Survey of the data stream clustering algorithms [20] An advantage over several areas such as market analysis, crime detection, also, this paper presents a survey of the usually-employed empirical methodologies. The study describes a summary of the data stream clustering problems. Dataset Used- GPS data from smartphones, web clickstream data.

III. PROPOSED METHODOLOGY

The above figure 1 shows entire proposed system execution using machine learning and classification approach. Initially system deals with large scale document data set, the basic assumption for this system to categories the data into 10-fold cross validation.

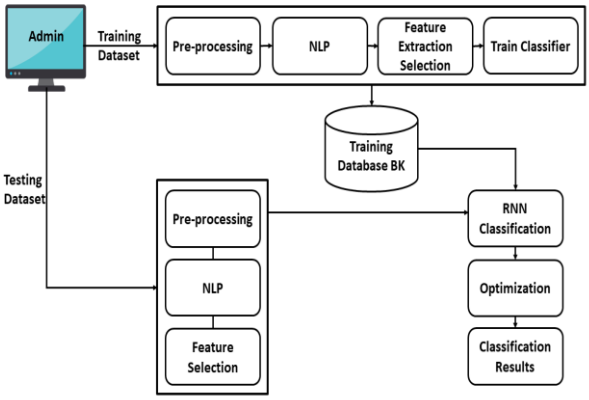

Fig. 1.Proposed System Architecture

In training phase, we extract some NLP based features and select using density-based approach [11]. The density-based approach is similar like TF-IDF. Once the feature selection has done, we build train model with the help of train classifier. It generates background knowledge for respect domains and store in knowledge base repository. These policies can be used for any kind of classification as well as clustering algorithm during the test phase. Propose system describes deep learning based classification algorithm as RNN. Three different activation functions have been used in given algorithm like sigmoid, Tanh, linear. Single input layer and single hidden layer have been used during the execution one feedback player has concurrently executed when system gets relevant feature set according to desired thresholds. The default Optimizations algorithm ADAM is available in deep learning library. Our hash-based strategy will reduce the optimization time complexity and provides flexibility to the system.

The consecutive execution of the four modules shows below.

A. Training and pre-processing module
B. Testing, preprocessing and TF-IDF module
C. Clustering module
D. Micro-clustering module

A. Training and pre-processing

- Uploading train dataset, the method reads the abstract part from PDF.
- Tokenizing, stop word eliminating, and Porter stemming.
- TF-IDF gives the accessibility of the current vector and reserve into the repository (DB).
- Training completed.
- BK completed for all classes.

B. Testing, preprocessing and TF-IDF

- Uploading the test dataset.
- TF-IDF computation used to recognize the weight of the current test instances.
- Extracting features with RNN and estimate a similar vector with all train features.

C. Clustering with Recurrent Neural Network

- Similar vector reflects the current density of the test objects with all training objects.
- Density-based clustering completed.
- Arranging the label according to the highest weight produced.

D. Micro clustering

- Micro clustering-based classification.
- Providing sub class categorization.
- Categorizing a particular cluster within various identical clusters from the main cluster.
- Classifying similarity scores for each bucket into the particular classes.

E. Algorithm TF-IDF

Input: Vectors V [i……n], every word from Vector as Term T

Output: Weight of TF IDF for each T

1. Vector = {doc1, doc2, doc3… doc-n}
2. For every comment.
3. Doc = {comment1, comment2, comment3, comment-n} and comments accessible in every document
4. Compute TF count as
5. TF (term, doc) = (term, doc)
6. term=particular term
7. doc= particular document in a term
8. IDF = term → sum(doc)
9. Return TF*IDF

F. Algorithm Recurrent Neural Network

- **Input:** Test_DB[], stored different test objects, Train_DB[], contains BK knowledge, and desired threshold Th for validating the current weight.
- **Output:** Hash_Map <class_label, sim_weight> all object's weight which crossed the threshold value**.**

1. Read every test object using function below

$$testFeature(m) = \sum_{m=1}^{n} (. featureSet[A[i] \dots \dots . . A[n] \leftarrow TestDBLits) \quad (1)$$

2. Eliminate every feature or input neuron from $testFeature(m)$ using the following equation.

$$Eliminated\_FeatureSetx = \sum_{x=1}^{n} (t) \leftarrow testFeature\ (m) \quad (2)$$

Eliminated_FeatureSetx[t] holds the feature vector of the particular domain.

3. Extract every train object using function below.

$$trainFeature(m) = \sum_{m=1}^{n} (. featureSet[A[i] \dots \dots . . A[n] \leftarrow TrainDBList) \quad (3)$$

4. Extract features from each test set as best features for specific document object $testFeature(m)$ using below function.

$$Eliminated\_FeatureSetx[t \dots . n] = \sum_{x=1}^{n} (t) \leftarrow testFeature\ (m) \quad (4)$$

Eliminated_FeatureSetx[t] holds the feature vector of the particular class/domain.

5. Evaluate every test vector with entire train features and generate a weight for respective instance.

$$weight = calcSim\ (FeatureSetx \parallel \sum_{t=1}^{n} FeatureSety[y]) \quad (5)$$

6. Return object [label] [weight]

G. Software and Hardware Requirement

- OS - Windows 7/8 Higher
- Programming Language - JAVA
- Tools– NetBeans, JDK 1.8 or Advanced
- Database - HeidiSQL 5.1
- Processor – i3 2.7 GHz.
- RAM - 4 GB
- Hard Disk - 300 GB
- Mouse - Logitech.
- Monitor - 15 VGA Color.

IV. RESULT AND DISCUSSION

The proposed system has implemented in two different ways first machine learning algorithm and other deep learning algorithms. In the existing system, we introduced experimental analysis with KDDCUP99 dataset, where a planned system introduced with deep learning approach with PDF document dataset. The system has evaluated according to classification accuracy and time complexity in the same environment. Figure 2 illustrates the classification accuracy for KDDCUP data set using a density-based technique [11]. Figure 3 carried out classification and clustering accuracy of the proposed system using PDF data set with RNN classification technique. Figure 4 shows the comparative analysis between different classifiers.

A. Existing System Results

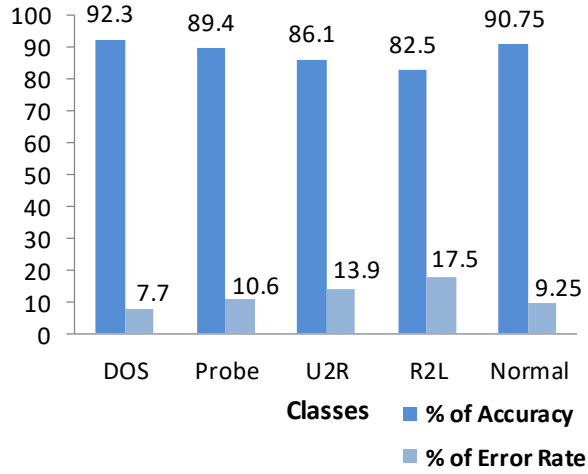

Fig. 2. Clustering Accuracy for KDD CUP99 Dataset Using ML(density-based)

The above figure 2[11] shows clustering accuracy kddCup 99 data, with five different classes. The average accuracy of the system using machine learning algorithm is around 87.50% for all classes.

B. Proposed System Result

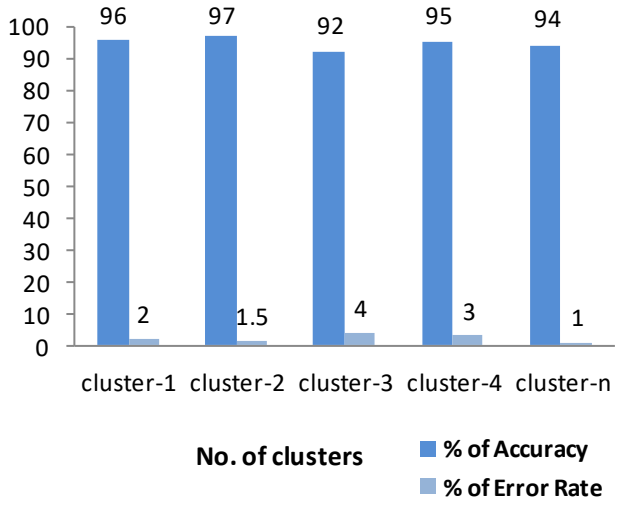

Fig. 3. Clustering accuracy for PDF dataset using deep learning (RNN)

Figure 3 shows clustering accuracy PDF dataset, with (n) different classes. The average accuracy of the system using machine learning algorithm is around 95% for all (n) classes.

C. Data Set

Partial implementation of the training program has around 100 document files for training.

Table-I: Training dataset

SR. NO	NO. OF PAPERS	DOMAIN NAME	LABEL
1	15	Data Mining	1
2	15	Machine Learning	2
3	15	Cloud Computing	3
4	15	Network Security	4
5	15	Soft computing	5
6	15	Artificial Intelligence	6
7	10	Image Processing	7

Table I show the details information about the dataset, which considered different classes/domains with assigned labels. The performance evaluation of the system will calculate the confusion matrix for the accuracy of this system.

D. Classification Performance

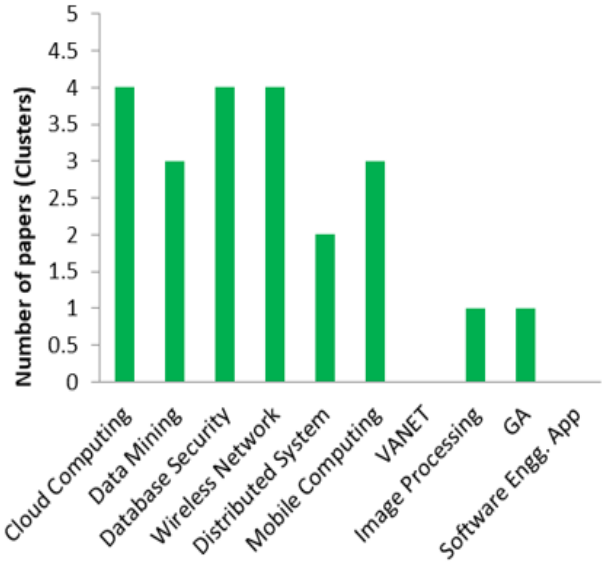

Fig. 4. Document Classification Performance

Fig 4 shows the Document classification result of the testing environment. We used a 30 % dataset for testing. Here x-axis indicates the different flavor of domains and the y-axis defines the number of papers (Clusters). The final result is based on a micro-clustering approach where sub-clusters are formed from the master cluster.

E. Processing Time

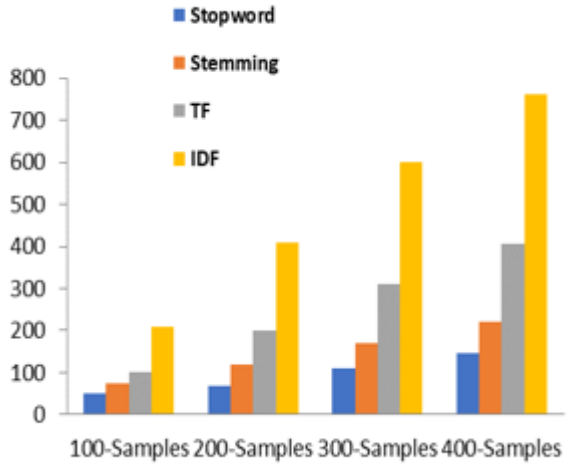

Fig. 5. System Processing Time

The above diagram shows the time required by each sample for each process, given the time is taken for the desired system configuration for a few seconds.

F. Comparison Between Different ML Classifiers

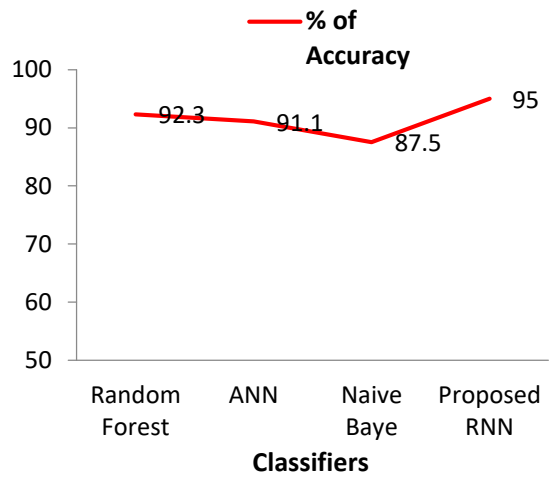

Fig. 6. Comparative analysis of different classifier algorithms

Fig. 6 Displays the performance analysis of the proposed system with a different flavor of current classification algorithms. according to RF, NB [20], ANN [21] statistics the RNN framework provides better reliability for the identification of structured data as it gives a minimum failure rate for the whole data set during the classification.

V. CONCLUSION

The proposed system illustrates the new document classification method based on Machine as well as Deep Learning (DL) with a micro-clustering approach. The micro clustering technique introduces to eliminate multi object classification issues. System initially proposed NLP feature extraction and selection approach and evaluate all those features using respective train classifiers. Recurrent neural network (RNN) provides highest accuracy for object classification, clustering, micro clustering for the heterogeneous data sets. In this research we used standard pdf data set to generate a cluster set for unlabeled data. Our algorithm introduces very less time complexity with higher accuracy around 95% for classification. Once the classification has done it categories all results into the separate clusters for micro clustering. Finally, we conclude the system can handle heterogeneous data set like structure, semi-structured, unstructured respectively. To deals with large unstructured data set into the distributed environment will be an interesting future enhancement for such systems.

REFERENCES

1. Medrouk L, Pappa A. Do Deep Networks Really Need Complex Modules for Multilingual Sentiment Polarity Detection and Domain Classification? In 2018 International Joint Conference on Neural Networks (IJCNN) 2018 Jul 8 (pp. 1-6). IEEE.
2. Kim CH, Kabanga EK, Kang SJ. Classifying malware using convolutional gated neural network. In2018 20th International Conference on Advanced Communication Technology (ICACT) 2018 Feb 11 (pp. 40-44). IEEE.
3. Jurca R, Cioara T, Anghel I, Antal M, Pop C, Moldovan D. Activities of Daily Living Classification using Recurrent Neural Networks. In2018 17th RoEduNet Conference: Networking in Education and Research (RoEduNet) 2018 Sep 6 (pp. 1-4). IEEE.
4. Jithesh V, Sagayaraj MJ, Srinivasa KG. LSTM recurrent neural networks for high resolution range profile-based radar target classification. In Computational Intelligence & Communication Technology (CICT), 2017 3rd International Conference on 2017 Feb 9 IEEE.
5. Hong D, Zhang Z, Xu X. Automatic modulation classification using recurrent neural networks. In Computer and Communications (ICCC), 2017 3rd IEEE International Conference on 2017 Dec 13 (pp. 695-700). IEEE.
6. Alom MZ, Alam M, Taha TM, Object recognition using cellular simultaneous recurrent networks and convolutional neural network. In2017 International Joint Conference on Neural Networks (IJCNN) 2017 May 14 (pp. 2873-2880). IEEE.
7. Abroyan N. Convolutional and recurrent neural networks for real-time data classification. In Innovative Computing Technology (INTECH), 2017 Seventh International Conference on 2017 Aug 16 (pp. 42-45). IEEE.
8. Kim J, Kim H. Classification performance using gated recurrent unit recurrent neural network on energy disaggregation. In Machine Learning and Cybernetics (ICMLC), 2016 International Conference on 2016 Jul 10 (Vol. 1, pp. 105-110). IEEE.
9. Zhang Y, Er MJ, Venkatesan R, Wang N, Pratama M. Sentiment classification using comprehensive attention recurrent models. In Neural Networks (IJCNN), 2016 International Joint Conference on 2016 Jul 24 (pp. 1562-1569). IEEE.
10. Salem A, Almarimi A, Andrejková G. Text Dissimilarities Predictions Using Convolutional Neural Networks and Clustering. In2018 World Symposium on Digital Intelligence for Systems and Machines (DISA) 2018 Aug 23 (pp. 343-347). IEEE.
11. Hahsler, Michael, and Matthew Bolaños. "Clustering data streams based on shared density between micro-clusters." IEEE Transactions on Knowledge and Data Engineering 28.6 (2016):1449-1461
12. Carnein, Matthias, and Heike Trautmann."EvoStream–Evolutionary Stream Clustering Utilizing Idle Times." Big Data Research (2018)
13. Mansalis, Stratos, et al. "An evaluation of data stream clustering algorithms." Statistical Analysis and Data Mining: The ASA Data Science Journal 11.4 (2018): 167-187
14. Fahy, Conor, Shengxiang Yang, and Mario Gongora. "Ant colony stream clustering: A fast density clustering algorithm for dynamic data streams." IEEE Transactions on Cybernetics (2018)
15. Tari, Zahir, et al. "MicroGRID: An Accurate and Efficient Real-Time Stream Data Clustering with Noise." Pacific-Asia Conference on Knowledge Discovery and Data Mining. Springer, Cham, 2018
16. Kumar, k. Naveen, raghu kumar, and g. Sreenivas. "Appraise on Various Clustering Modules of Clustering Data Streams based on Shared Density between Micro-Clusters". (2018)
17. Carnein, Matthias, Dennis Assenmacher, and Heike Trautmann. "An empirical comparison of stream clustering algorithms." Proceedings of the Computing Frontiers Conference. ACM, (2017)
18. Hahsler, Michael, Matthew Bolanos, and John Forrest. "Introduction to stream: An Extensible Framework for Data Stream Clustering Research with R." Journal of Statistical Software 76.14 (2017)
19. Desai, Prashant V., and Vilas S. Gaikawad. "Novel approach for data stream clustering through micro-clusters shared Density." (2017)
20. Kulathunga, Chalitha, and D. D. Karunaratne." An ontology-based and domain-specific clustering methodology for financial documents", advances in ICT for Emerging Regions (ICTER), 2017 Seventeenth International Conference on. IEEE, (2017)
21. Pacifico LD, Macario V, Oliveira JF. Plant Classification Using Artificial Neural Networks. In2018 International Joint Conference on Neural Networks (IJCNN) 2018 Jul 8 (pp. 1-6). IEEE
22. Kamran, Kiana, Mujtaba, Sanjana "Text Classification Algorithms: A Survey" MDPI,17 April 2019; Published: 23 April 2019

AUTHORS PROFILE

Amirkhan Mulla was born in Karad, MH India, in 1992. He is working in the IT industry since 2016 as Oracle Database Administrator. He did MTech in Computer Science and Engineering in Rajarambapu Institute of Technology(an autonomous institute), B.E. in Computer Science and Engineering, Diploma in Computer Engineering from the Shivaji University, Kolhapur, respectively. His research interests include Machine Learning, Data Mining, Database Operations, and Engineering.

Retrieval Number: B3716129219/2019©BEIESP
DOI: 10.35940/ijeat.B3716.129219
Journal Website: www.ijeat.org

Published By:
Blue Eyes Intelligence Engineering
& Sciences Publication

Sachin Patil was born in Mumbai, India, in 1981. He received the B.E. degree in computer science and engineering and MTech in computer science and technology from the Shivaji University, Kolhapur in 2003 and 2011 respectively. He is pursuing a Ph.D. degree in computer science and engineering under A.I.C.T.E. Q.I.P. scheme at Walchand College of Engineering (Govt. aided and an Autonomous Institute) affiliated to Shivaji University, Kolhapur, MH India. Since 2010, he has been an Assistant Professor in the Computer Science and Engineering Department, Rajarambapu Institute of Technology, Rajaramnagar, MH – India. He has worked as head of Computer Science and Engineering department at Rajarambapu Institute of Technology, Rajaramnagar, MH – India. He is the author of two book chapters viz. Springer-Verlag and Taylor and Francis. He has more than 15 research papers. His research interests include Database Engineering and Big Data analytics. He has received a "Distinguished Facilitator" award at Inspire faculty contest organized by Infosys, Pune.