JETIR.ORG ISSN: 2349-5162 | ESTD Year : 2014 | Monthly Issue
JOURNAL OF EMERGING TECHNOLOGIES AND INNOVATIVE RESEARCH (JETIR)
An International Scholarly Open Access, Peer-reviewed, Refereed Journal

A SURVEY ON DATA MINING TASKS WITH MACHINE LEARNING TECHNIQUES, TOOLS, APPLICATIONS AND OPEN CHALLENGES

[1]K K Vinoth Kumar, [2]Dr. K P Lochanambal
[1]Research Scholar, PG and Research Department of Computer Science,
Government Arts College, Udumalpet, Tamil Nadu, India
E-mail: vinothphd1983@gmail.com
[2]Assistant Professor, PG & Research Department of Computer Science,
Government Arts College, Udumalpet, Tamil Nadu, India.
E-mail: kplmca@rediffmail.com

Abstract:

*D*ata mining techniques involved in extracting needful information from large amount of structured and unstructured data, and generate accurate and new set of data to any individuals or organizations; Data Mining techniques fails to make decisions because it retrieves only relevant information. To make machines decisions we using predictive automated advance technologies like machine learning techniques to improve the use of technologies in many real-time applications to improve the accuracy, efficiency in finding the solution by making right decision without human interruption to any kind of applications. In this paper we did survey on various tools, techniques, algorithms, applications using machine techniques based on size, feature, characteristics of the data, and nature of the application domain. This architectural framework will be very useful for future machine learning and Deep learning solutions to make decisions by learning different algorithms for different input data models.

Keywords: Data Mining, supervised Learning, Unsupervised Learning, Artificial Intelligence, Machine Learning, Deep Learning.

I. Introduction:

Data Mining refers as Knowledge Discovery in Database. It extracts relevant information from the underlying pattern from large databases [1], KDD process involved in cleaning irrelevant data from the database through implementing different mining algorithms and sources from different databases are integrated, and based on the task relevant data retrieved from database are transformed data for mining process to aggregate various operations.

[2], In order to obtain relevant information data mining algorithms applied after aggregation, finally using some measurements data patterns are identified and represented to end-users by applying representation technique on data patterns. all these set of process done by cleaning Irrelevant data's, integrating different related data sources, selecting appropriate data, transforming relevant data for mining, extracting data using mining methods, evaluating patterns and finally representing knowledge.

Data mining tasks influences by other learning techniques, such as artificial intelligence, machine learning, deep learning and data analytical experts works as data scientists, and mining algorithm mainly focus on measuring and calculating the instances and features using algorithms.

Architectural diagram clearly points input task given to the machine is either predictive, descriptive or optimization tasks, all the given mining techniques such as, clustering, classification, association, regression and summarization, all these listed techniques are either learned by supervised, unsupervised or reinforcement learners.

- **1.1. Descriptive:** This term is generally used to find data similarities from existing patterns, unstructured data is grouped and transformed into meaningful information
- **1.2. Predictive:** predictive analysis is developed and designed to analyze the past set of data patterns and to accurately predict the new data set.
- **1.3. Optimization:** Method uses historical data and estimate outcomes based on variables and offers best successful suggestions about outcomes by using machine learning algorithms.
- **1.4. Clustering**: Unsupervised Learning (Unlabeled data), based on the similarity on the data will group the data together and mathematical concepts are used to measure the similarity distance between data points.
- **1.5. Association:** This mining technique finds most continuously occurring data items or inter-related elements.

1.6. Summarization: Relevant data set is summarized from different resources and result reports aggregates as a smaller set of data.

1.7. Classification: It is a supervised learning (Labeled data), past labeled data has been used and already we know the output. To predict two classes we using classification problem, another name is binary classification (Results in 0 0r 1). If we want to classify more than two classes, then it is said to be a multi-classification (Prediction is discrete).

1.8. Regression: A target value is modeling by independent predictors. Regression method mostly used to find out prominent relationship among variables, regression method varies based on the individual variables and type of relationship between dependent and independent variables.

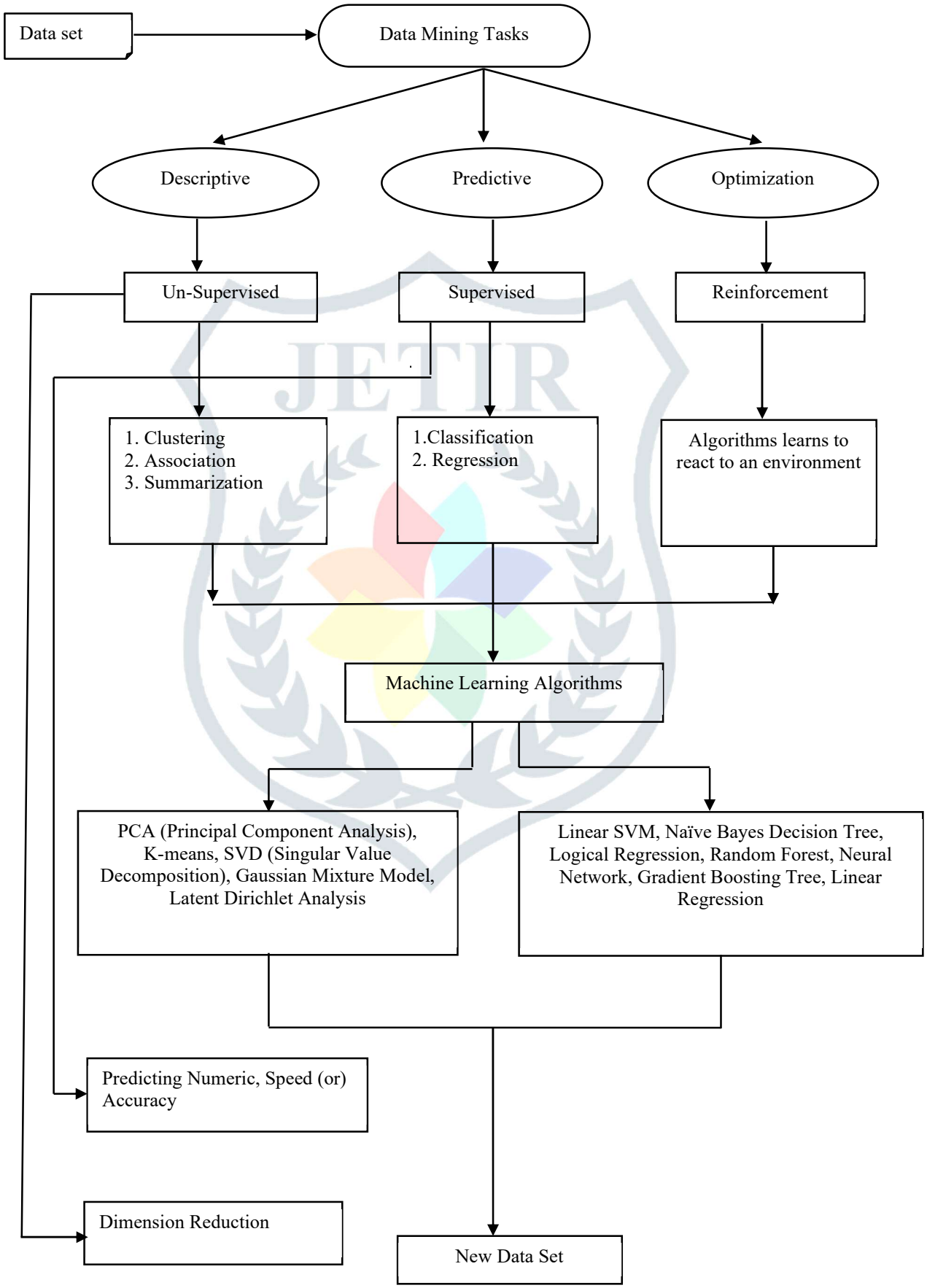

Fig.1. Architectural Diagram for Data Mining Tasks and Machine Learning Algorithms

1.9. Reinforcement (or) Time Series Analysis: Algorithm learns through feedback provides by past experiences, each step of desired feedback from previous outcome is next step for algorithm to reach goal set, and decides the next step for further process and uses trial and error approach to achieve goal set.

Some of the, machine learning algorithms is, K- Means Clustering, Principal Component-Analysis, Fuzzy K-Means, Hierarchical Clustering, Spectral Clustering, Neural Networks, Naïve Bayes, Decision tree, Logistic regression, Expectation Maximization and singular value decomposition. Let's discuss with some learning algorithms used in previous research articles.

[3], K-means clustering aiming in finding the closest cluster center from the given n-dimensional dataset. It eliminates the irrelevant features but measuring the similarity. Mini batch K++ clustering is more efficient compare to K-means clustering, because it avoids the local optimum results by using K++ clustering algorithm.

Expectation Maximization and Gaussian Mixture [4] (EM & GM), it uses to identify the maximum likelihood (MLE) and variables of particular parameter are estimated. [5] Gaussian Mixture is responsible for likely or close points to the currently existing cluster center by finding neighborhoods and classifies most prominent class labels.

[6], Principal Component Analysis (PCA)-It deals with dimensionality reduction, exactly eliminates the redundant data from the given input features or images by transforming the features into the column and row vectors to find the average mean value of the input vectors. PCA is subject to find the Orthonormal vector, Eigen vectors and Eigen values of the Covariance Matrix.

[7], FCM-Fuzzy c-means and K- means are efficient methods commonly used in analyzing clustering algorithms. Objective of FCM and K-means is similar way in finding the dissimilarity between the data points and optimally clusters, and partitioning the attributes and variables of the objects into homogeneous group.

Ensemble Clusters- [8], Deals with the great idea of integrating the multiple weal learners to strengthen the model in effective manner for better outcome of results, and it deals with the datasets in imbalanced state. Both boosting and bagging techniques executes more accurate results by aggregating the multiple learners.

[9], Random Forest-For multiple decisions making independent trees are combined by the random forest ensemble method, here training data of tree is constructed using the bootstrap samples, for every individual tree feature selection performed and subset created for feature space. Prediction of random forest is concluded by assigning the majority vote.

[10], Decision Tree-Familiar and well-known classification method, it is simple and easy of interpretability on classification rule: here decision function is deals with non-leaf node and the feature vector function is binary valued, every leaf node is associated with class label.

Support Vector Machine (SVM)- [11], SVM is a most popular classification method classifies by separating the hyperplane, it is a binary classification method, output will be either o or 1 based on the threshold values achieves best performance in classification.

II. Related Works:

Learning Techniques are categorized into supervised, semi supervised, unsupervised learning and reinforcement, all these listed techniques are performing statistical and mathematical approaches on machine and analyze the input data. Artificial Intelligence enables machine to think and take decisions without any human intervention; car without human is the best example for the AI application, and it derives by using [12], Machine learning and deep learning Algorithms. Data science (analyze, model and report) is work on artificial intelligence, machine learning and deep learning to perform statistical, probability and algebraic calculations.

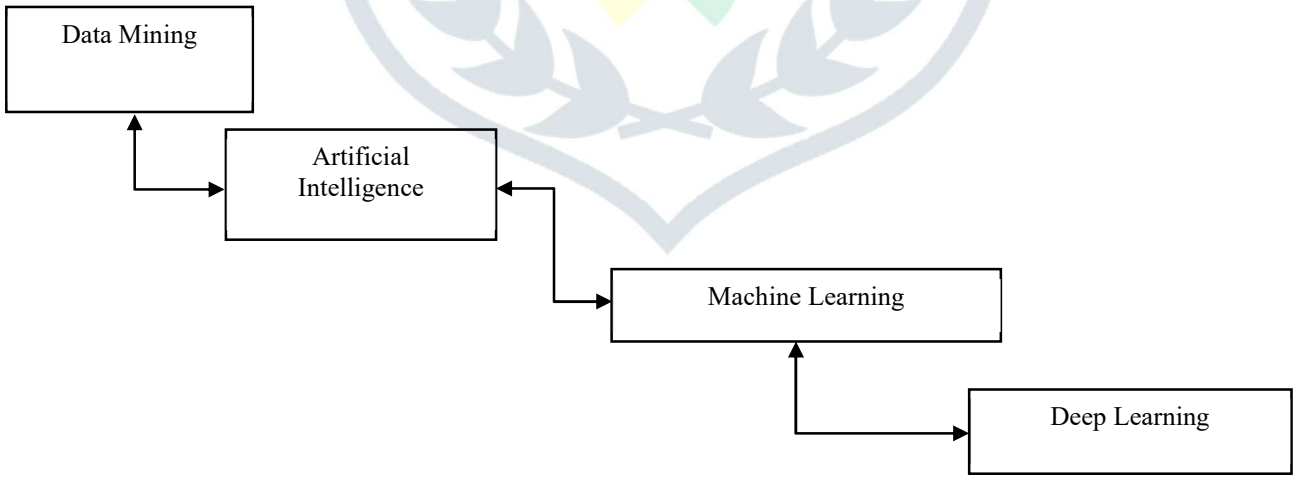

Fig.2. Hierarchical representation of learning techniques

In above diagram clarifies about the interrelationship between the learning algorithms. Deep learning is the subset of the machine learning and machine learning is the subset of artificial intelligence. These three-learning method influences the data mining techniques.

III. Deep learning Techniques:

Modeling idea of [13], deep learning resembles the thinking concept of human brain. The most improved thing in deep learning is multi-neural network architecture.

3.1 Classic Neural Network:

This type of fully connected Neural Network is mostly works for table datasets which contains rows and columns in CSV format and real value as an input.

Three functions used in CNN are:

i) Sigmoid Curve
ii) Hyperbolic Tangent (Tanh)
iii) Rectifies Linear Unit (ReLU)

3.2 Convolution Neural Network:

[14], CNN is special advanced efficient artificial neural network model. Its ability to solve highly complex problems like image segmentation, image processing, video analysis and natural Language processing. One of the efficient models with best outcome for given input through passing several hidden layers.

CNN has Four Sequence of Process Stage

i) Convolution
ii) Max-Polling
iii) Flattering
iv) Full Connection

3.3 Recurrent Neural Network:

It designed for the purpose of image classification. [15] Sentiment analysis problems and single image includes many number of words like image captioning.

RNN analyzing problem into two different approaches:

i) Long short-term memory ($LSTM_s$)
ii) Recurrent Neural Network Cheatsheet (RNN_s)

3.4 Generative Adversarial Networks:

[16], GAN widely used in finding real and false image processing in different applications, and also discovers new drug discovery process, text and image processing. GAN hybrids two deep learning process.

i) Generation
ii) Discrimination

3.5 Self-Organizing Maps:

The [17], SOM is plays important role in reducing the dimension size of data model, it mostly used to develop artificial intelligence related text processing, video processing, audio processing projects and analyzing frameworks of real time datasets in projects.

3.6 Boltzmann Machine:

This type of neural network designed nodes in the form of circular arrangements. This is stochastic model compare to all other neural network models. [18], BM-Used to analyzing some specific set of datasets, used in binary representation domain-based platform and system monitoring.

3.7 Deep Reinforcement Learning:

DRL-is an agent-based network activity model to achieve the objective by interacting different situations. This type of deep learning network model has several types of hidden layers between input and output layers to predict the forthcoming situation, DRL mostly used in automatic car driving, poke, chess games, asset pricing in finance and inventory management.

3.8 Auto Encoders:

[19], Most of the common deep learning technique, special activity of this technique automatically operates the model based on the input given to the machine or model. Most of the leveraging inherent data structure uses this technique, autoencoders used for feature detection and add extra features nth the large datasets. Four types of Auto encoders:

i) Sparse
ii) Denoising
iii) Contractive
iv) Stacked

3.9 Back propagation:

[20], Technique widely used in data debugging, initially it works on parameters based on network analysis the data, and loss function is used to weighted out data. Finally, errors find and automatically adjusted by model.

3.10 Gradient Descent:

Logically updates parameters simultaneously in the given model. [21], Gradient is a mathematical term refers slop; it measures the similarity and dissimilarity between the variables, mostly used in finding the optimum solution. Gradient descent uses convex function to follow and record the data flow rate.

IV. Data Set partitioning:

If we started to train a machine learning model, we need to split data set for training and testing. For testing collecting separate data is highly expensive and time-consuming.so we splitting the existing structured data into different sections, one is for training (70% of labeled data), remaining data is for testing (30%). Training and testing model should have similar accuracy for

better model, under fitting and over fitting of the model will be apparent in testing section. In machine learning training and testing commonly used for evaluation

V. Validation:

To check the performance of the predicted model from training data, we involved in testing process refer as cross validation and it measures the efficiency and the performance of model. There is variety of cross validation methods used for performance evaluation.

VI. Efficient Tools for Machine Learning:

In application development area most of the Data analytics are prioritizing for python and secondary priority for R language, because python supports collections of extensive special libraries compare to all other languages, easy to use and performing machine learning algorithms faster compare to other languages. To solve many scientific computations NUMBY-Library used for python. Machine learning in python mostly uses Pybrain Library for efficient outcome and best performance. In Bioengineering and bio-informatics field R is mainly used for statistical analysis and visualization. Most of the organizations using Java for secured network in fraud detecting algorithms.

VII. Machine Learning trending field and its applications:

As world is emerging and adapted with Artificial Intelligence in any kind of domain. To fulfill the requirement of business and professional organizations simultaneously technology rapidly developing in advance, widely used popular applications are interacting and gained attention through deep learning, machine learning through artificial intelligence. Some algorithms and statistical calculations make ability to understand the data models and problems, most of the researchers and scientists understanding the technology and implementing the data model technically based on the requirement. [22], Application interacts and integrates with machine learning techniques make developer and researcher to think and highly understand to handle the algorithm with newly introduced tools and techniques for better outcome of accuracy.

Machine learning techniques plays a vital role in many trending fields, such as healthcare predictive analysis, professors, retail business, image processing, data scientist, business analyst, spatial data analytics, architectural analysts and especially research scholars implementing their idea using machine learning techniques.

VIII. Conclusion:

This survey presented a variety of algorithms applied to machine learning, but commonly all are technically performing mathematical and statistical calculations. Every individual machine learning algorithm has its own strength and weakness based on size, nature, characteristics of data and applications used. If the dataset is larger some algorithms will perform well and for smaller vice versa, complexity of data chooses and decides the machine learning algorithms otherwise over fitting will result in bad performance of algorithms. Outcome of the accuracy, performance and execution will be vary based on tools, techniques, algorithms, program languages and applications used. Machine learning and deep learning techniques are the subset of artificial intelligence, all these techniques influences data mining techniques; finally, data science is working on Data Mining, artificial intelligence, machine learning and deep learning to perform modeling, analyzing and reporting.

References:

[1] Akshat Savaliya, Aakash Bhatia, Jitendra Bhatia, "Applications of Data Mining Techniques in IoT: A Short Review", International journal of Scientific Research in science, Engineering and Technology", Volume 4, Issue 2, 2018, Pp.218-222.

[2] Hanan Abdullah Mengash, "Using Data Mining Techniques to Predict Student Performance to Support Decision Making in University Admission Systems", IEEE Access, Volume 8, March 2020, Pp.55462-55470.

[3] Kai Peng, Victor C.M.Leuang , Qingjia , Huang , "Clustering Approach Based on Mini Batch K-means for Intrusion Detection System over Big Data" , IEEE Access , 2018 , Pp.1-9.

[4] Mayank Tyagi, Francesca Bovolo, Ankit.K.Kehra , Subhasis Chaudhuri , and Lorenzo Bruzzone , " A context-Sensitive Cluster Technique Based on Grapg-Cut Initialization and Expectation-Maximization Algorithm " , IEEE Geoscience and Remote Sensing Letters , Vol.5 , Issue.1 , January 2008 , Pp.21-25.

[5] Min Wag, Sherief Abdelfattah, Nour Mustafa, and Jiankun Hu, "Deep Gaussian Mixture –Hidden Markov Model for Classifcation of ECG Signals", IEEE Transactions on Emerging Topics in Computational Intelligence, Vol.2, Issue.4, August 2018, Pp.278-287.

[6] Sekhar Rajandran, amit Kaul, Ravinder Nath, A.S.Arora and Sushil Chauhan , "Comparision of PCA & 2D PCA on Indian Faces" , International Conference on Signal Propagation and Computer Technology , IEEE , 2014 , Pp.561- 566.

[7] Miin-Shan Yang and Chih-Ying Lin, "Block Fuzzy K-Modes Clustering Algorithm", IEEE, 2009, Pp.384-389.

[8] Dong Huang, Chang-Dong Wang, and Jian-Huang Lai, "Locally weighted Ensemble Clustering", IEEE Transactions on Cybernetics, 2017, Pp.1-14.

[9] Zhi-Sen Wei, Jing-Yu, Hong-Bin Shez and Dong- Jun Ya, "A Cascade Random Forests Algorithm for Predecting Protein-Protein Interacting Sites", IEEE Transactions on Nano Bioscience, IEEE, 2015, Pp.1-31.

[10] Naresh Manwani, and P.S. Sastry, "Geometric Decision Tree", IEEE Transactions on System, Man, and Cybernetics, Volume.42, Issue.2, Feburary 2012, Pp.181-192.

[11] Yan Yang, Juan Wang, and Yongyi Yang, "Improving SVM Classifier with Prior Knowledge in Microclassification Detection", IEEE, 2012, Pp.2837-2.

[12] L. Silva, L.Utimura , K.Coasta, M.Silva and S. Prado ,"Study on machine Learning Techniques for Botnet Detection" , IEEE Latin America Transactions , Volume 18 , Issue 5 , May 2020 , Pp.881-888.

[13] Ling Wu, Qishan Zhang, Chi-Hua Chen, Kun Guo and Deqin Wang, "Deep learning Techniques for Community Detection in Social Networks", IEEE Access: Special Section on Data Mining for Internet of Things, Volume 8, June 2020, Pp.96016-96026.

[14] Guangdong Song, Jiulong Cheng, and Keneeth T.V. Grattan, "Recognition of Microseismic and Blasting Signals in Mines Based on Convolutional Neural Network and Stock Well Transform", IEEE Access: Special Section on Machine Learning Designs, Implementations and Techniques, Volume 8, March 2020, Pp.45523-45530.
[15] Hanyong Shao, "Delay-Dependent Stability for Recurrent Neural Network with Time-Variying Delays", IEEE Transactions on Neural Networks, Volume 19, Issue 9, September 2008, Pp.1647-1651.
[16] Chaoyue Wang, Chang Xu, Xin Yao, Dacheng Tao, "Evolutionary generative Adversarial Networks" IEEE Transactions on EvolutionaryComputation, Draft, Accepted 2019, Pp.1-14.
[17] Xiaofei Qu, Lin yang, Kai Guo, Meng Sun, Linru Ma, Tao Feng, Shuangyin Ren, Kechao Li, and Xin ma, "Direct Batch Growth Hierarchical Self-Organizing Mapped Based on Statisticsfor Efficient Network Intrusion Detection" IEEE Access: Special Section on Emerging Approaches to Cyber Security, Volume 8, March 2020, Pp.42251-42260.
[18] Wei Huang and Huimin Yu, "Cooperative Object Segmentation and Recognition via Restricted and BoltzmannMachine" Electronics Letters, Volume 56, Issue 8, April 2020, Pp.378-380.
[19] Nicholas Merrill, Azim Eskandarian, "Modified Auto Encoder Training and Scoring for Robust UnsupervisedAnomaly Detection in Deep Learning", IEEE Access, Current Version May 2020, Pp.1-11.
[20] Olga Krestinskaya, Khaled Nabil Salama, and Alex Pappachen James, "Learning in Memristive Neural Network Architecuturs Using Analog BackPropagation Circuits ", IEEE Transactions on Circuit and Systems, Pp.1-14.
[21] Yi Sun, Yan Tian, Yiping Xu, and Jianxiang Li, "Limited Gradient Decent: Learning with Noisy Labels" IEEE Access, Volume 7, December 2019, Pp.168296-168306.
[22] Zhenni Feng and Yanmin Zhu, "A Survey on Trajectory Data Mining: Techniques and Applications", Special sections on Theoretical Foundations for Big Data Applications: Challenges and Opportunities Volume 4, May 2016, Pp.2056-2067.