Comparative analysis of Unsupervised Machine Learning Approaches

Navjot Kaur, Harkiran Kaur
Department of Computer Science and Engineering
Thapar Institute of engineering and technology(Deemed to be University), Patiala, India
navrndhawa29@gmail.com, harkiran.kaur@thapar.edu

Abstract

Unsupervised machine learning is a very wide domain of interest. This field is further used in a number of fields which is capable of solving numerous worlds, which cannot have predictable knowledge. This paper shows recompense of many approaches to unsupervised machine learning with evaluation of existing approaches' in a different domain of interest. The result of this provisional testing is identification of the unsupervised machine learning approaches which are superior to existing approaches for resolving the crisis that arises in different areas of interest.

Keywords: **Unsupervised machine learning, K-means, Expectation maximization, Scenes classification.**

I. INTRODUCTION

Unsupervised machine learning is machine learning mission of inferring a gathering hidden structure from unlabeled data; which is not constantly obtainable. This concept is used in various types of classifiers and algorithms. It discovers a function to describe hidden structure from unlabeled data, which offer unlabeled examples to learner. Unsupervised learning essentially used when data is not labeled, means in unsupervised learning, first of all perform clustering on similar type unlabeled data and later perform any other operations. This is used in many areas like signal detection, pattern recognition, adaptive system and any other systems. Different approaches used for unsupervised machine learning like clustering, neural network. Many research works have been undertaken on unsupervised machine learning in different areas. Pattern recognition aspect focuses on reliability of different patterns in given data using different algorithms of unsupervised machine learning. Various applications of pattern recognition include; Classification of spoken words and classification of hand printed characters. With the help of unsupervised adaptive algorithm, spoken words and hand printed characters can easily classify. Adaptive algorithm used linear classifier for this reason. In this pattern recognition, Digital image processing helps to analyze accurate image from complex images containing multiple things. Two types of Pattern recognition approaches are: Sequential decision approach and Bayesian decision approach. Adaptive system uses linear approach, in which observations of previously processed samples have been, used which helps to improve performance of classifiers. These systems are called as decision directed learning schemes.

II. LITERATURE REVIEW

Conventional classifiers used human operator for trait extractions, for instance: histogram of oriented gradients (HOG). These methods are very composite and time consuming for any other category of data such as audio, video and images. Trait extraction from dataset and guidance of classifiers are main inconvenience of classification. These difficulties can be removed by using new approach called unsupervised feature learning-extreme learning machine (UFL-ELM). This approach is widely and rapidly applied on universal data and large datasets. When this works on experiments of large dataset of images, these experiments trained this technique by authenticating the usage and speed of the technique. It is feed forward single hidden neural network. This produced two results, one eliminate the requirements of hand craft feature and, another increases the training

speed of classifier. For increasing the training speed of the classifier, considerable parallel programming used, which implemented by using CUDA library. These results are proved by using 60,000 color images of 32*32 dimensions in 10 classes [1].

For a network traffic classification, earlier established methods like port based internet traffic classification and payload internet traffic classification has been used. In a port based technique port numbers are used for classification. Some difficulties occur in this technique, one of them is the occurrence of unpredictable port number in some applications and other is the less accuracy. This accuracy supplementary improved nearly 79% in payload technique. This technique is based on characteristics signatures of known application. At the present time, network traffic can be classified by using unsupervised machine algorithms K-means and Expectation Maximization algorithms. These algorithms enhance the accuracy of network traffic classification. This proved that the accuracy of K-means algorithm is more than the Expectation Maximization algorithm. Experimentally, when consider an 80 number of clusters, accuracy of K-means is 88% and Expectation Maximization is 84% [11].

The previous studies shows that the Content based image retrieval CBIR are unsuccessful in reducing the gap between human semantics and images of low level content. Content based image retrieval is technique used for image retrieval crisis. These semantic gaps can be reduced by using unsupervised machine learning. CBIR uses unsupervised machine learning for creating bridge between high level semantic and low level visual descriptor. Also, the lack of standard image database availability and unified performance measures cannot occur in CBIR [12].

For the interaction of robot with environment, there are some good model such as Piece Wise Smooth- Hybrid System (PWS-HS) that are used for identification and tracking. This model uses many existing techniques for this purpose such as Single GP, Switching GPs. These techniques fail to capture the multimodality of subsystem near to mode transition. Usage of unsupervised learning framework for PWS-HS, remove composite robot system problems. Two challenging problems included the unidentified domains of subsystems and the identification of mode transition circumstance. Many algorithms are used, which are able to learn the hybrid system as unsupervised framework. This algorithm gives the structure of hybrid system, which finds the domains of subsystems and learn mode transition, reset subsystems separately. Experimentally it is shown that unsupervised PWS-HS captures the multimodality of subsystem near the mode transition [6].

Evaluation of traffic state variation patterns of urban road network is very complicated. These patterns vary with different roads and social activities. For the extraction of these patterns, spectral clusertering technique is used, which is unsupervised learning method. This technique helps to analyze daily traffic variations of road network based on different sections based traffic speed datasets. This method does two works parallels, one is the extraction of traffic variation features and another is transfer of traditional clustering problems into graph partition problem. Traditional clustering problems extract the feature with single or less attributes; graph partition problems extract features with multiple attributes in high dimension space. Experimentally, consider 5 clusters for this purpose. Element of each clusters are heterogeneous and relates the frequency of distribution of sections for each cluster with road hierarchy and functions. This work is extended in future with traffic prediction and route guidance algorithms [10].

For 2D and 3D feature matching and object recognition, graph matching can be applied. Graph matching is a big challenge in the computer vision. Publishing of parameter learning helps to control graph matching and also, improves the matching rate. Unsupervised parameter learning improves performance of graph matching algorithms, in terms of efficiency and quality as compare to supervised parameter learning. Unsupervised parameter learning, helps in learning of parameters which are unknowns. This can be shown experimentally, how to unsupervised parameter learning improves the performance of graph matching algorithms [7].

Evaluation of motor rehabilitation process is a complex process and it is difficult to evaluate using handcraft feature based method. Mainly two problems are faced in this method, first is instruments used for this method is very expensive and need technical skills to use these instruments. Secondly, these methods are motion specific that is these methods evaluates features for specific movements. The work can be done easily by using unsupervised learning method for anomaly detection, learn normal movement's pattern. These normal movements pattern learnt from wearable inertial measurement unit such as gyroscope. It is more general than the handcraft feature method because this is directly from raw data without any specific feature. Auto encode, is a form of neural network with 3 layers including the input layer, hidden layer and output layer. Auto encodes

reconstruction error helps to detect the anomalous movement with high accuracy. Auto encode also allows the evaluated system to evaluate new types of movements with reasonable amount of time [9].

Mostly organization used cloud services for the file sharing. So they add many new channels for insider threats including personal identifiable information, company IPs, source code etc. For detection of these threats used two stage machine learning system, which can automatically detect these threats. In the first stage, development some access logs data onto two relationship graph such as user/user and user/file and apply 3 graphs based unsupervised learning algorithms: Oddball, Page Rank and local outlier factors (LOF). These algorithms help to generate outlier's indicators. In second stage, group the outlier indicators and introduce the discrete wavelet transform (DWT) method with Haar wavelet function to evaluate users' temporal behavior for detection of insider threats. The coefficient of first level of the DWT with Haar wavelet function is able to effectively capture the user behaviors temporal pattern. In the future work more data source can be used, including but unlimited to, active directory logs, to better detection of insider threats [13].

Object recognition method is used in many fields. Detection of live fish recognition is one of fisher survey, which has main challenge, underwater image recognition measurements. This measurement is done with low quality image, uncontrolled objects and also problem occurs during the representation of samples. Existing techniques of feature extraction for square measurements need human management. Now E.Steffi et.al develop algorithm or under water fish recognition framework using an unsupervised learning algorithm with non-rigid part model. This model is developed to find the meaningful part of fish body using object component which support relaxation labeling for matching the object component properly. This is also generating binary category hierarchy for a classifier. Experimental result of this model is shown that this is achieving high accuracy on each public land and finds the underwater image of fish with high uncertainty [2].

Gradually, increases the demand of wireless communication. So many allocated spectral resources utilized in a less amount. When primary user emulation attacks (PUEA) and spectrum sensing data falsification (SSDF) are present together in radio network, then there is problem in detection of secure sensing in this network. We can use secure sensing algorithm for this problem. This algorithm is identified the secondary users attack, when no information of identification error and secondary user (SU) given in advance. In this firstly find the identification error and different reliability of secondary user, assign identify value to these SUs and update SUs. This algorithm show better performance than the conventional secure sensing algorithms [14].

In a single video have many human actions such as in a video of road some people driving car, some walk along the road. Computer can not automatically extract these human actions from a video. So unsupervised methods can be used for extraction of these human actions from a given video. The unsupervised learning algorithms automatically learn all features of human actions. This is done by using latent topic models such as latent dirichlet allocation (LDA) and probabilistic latent semantic analysis (pLSA) model. These methods find and handle noisy features in a dynamic background or moving picture. The algorithms categorized, localize the human actions from a video. These algorithms test with 3 datasets: KTH human motion dataset, the Weizmann human action dataset, recent dataset of figure skating actions. This testing leads to future investigation [4].

New development and uses of internet increases, this creates very complex network environment. For this complex network needs security. For security purpose use intrusion detection, uncover unauthorized access to computer network. Previous methods of intrusion detection produce alert with large number of false positive. This increase the burden of network administration. When combined the K mean clustering, J48 Decision tree classification, latest self organizing map, then they remove number of false positive. K-means perform clustering fast and with a typical fashion. J48 Decision tree classification used to find best subset feature, which in future used for analyze the most false attack and reduce false positive rate. Latest self organizing map is used to find similarity between types of attack and improve accuracy. Result of this method; find countable reduction of false positive detection [5].

Growth of sequential data sets increases broadly with the development of storage technique and any other techniques. Sequential datasets used in many applications such as video motion analysis, speech recognition. For finding the structural group in unlabeled dataset, the paper utilizes the unsupervised machine learning technique. Hidden Markov Model (HMM) is one of method of machine learning used for this purpose. Now, HMM is replaced with student's t-mixture mode (SMM), called Student's t-HMM (SHMM). Previously, SHMM with finite mixture model was applied for expectation of maximization algorithm, which was used to find the unknown parameters. Also it fails to find relevant feature within the local subspace and affected the

performance within clustering. Yuhui Zheng et.al overcomes this problem for developed a SHMM by combing SMM with LFS (localized feature saliency). It improves the performance within clustering. It is used to evaluate the important local feature. Also, it uses the variation Bayesian learning technique for estimating LFS, number of components and other parameters simultaneously. Result of this is the real data and synthetic dataset which finds the effectiveness and accuracy of our model [15].

For treatment of Traumatic brain injures (TBI), the intracranial pressure (ICP) technique is used. This technique monitors the pressure within brain or skull and this is very useful for treatment of brain problems. Nature of ICP is invasive, so measurement of this is done using invasive technique. In terms of accuracy this technique is good but it is time consuming. Parisa Naraei et.al proposed hybrid approach with two problems: One, Wavelet Analysis, useful for dividing waveforms into that different scale component, which can be viewed when represented arbitrary waveforms in time versus frequency representation for finding potential patterns. Second, K means clustering, unsupervised learning method of clustering, which does not need to learn data from previously labeled data. This approach, called wavelet based K mean clustering analysis, is used for the detection of the physiological signal patterns and study of changes, which is occurred in this technique. This is experimentally performed on twenty patients and results of this are useful and less time consuming method for the detection of pattern [8].

Now, satellite imaging sensor acquires a lot of high-resolution image datasets, face the problem to access these datasets directly. Many methods are given for scene classification; one is bag-of-visual-words (BoVW). This is divided into two parts: feature learning and feature encoding. This is a highly efficient method, but it doesn't give spatial and structural information in these datasets. Fan Zhang, proposed an unsupervised learning method for scenes classification when very high-resolution dataset are given. The unsupervised method is saliency detection algorithm. This is extracted unlabeled data patches from the salient region in given image data set and future used unsupervised feature learning method for feature extraction of sets which is efficient. This shows the statistics which are generated from this learned feature extraction method which is characterized by a complex scene. This works very well and gives good classification accuracy. For good Accuracy, this algorithm uses dropout technique. The Future work includes the extension of this method for learning hierarchy feature of an image with representation from low to high-level feature [3].

TABLE I. COMPARATIVE ANALYSIS OF UNSUPERVISED LEARNING TECHNIQUES

Parameters	Area of Concern	Existing Techniques	Latest Techniques
Time	Feature learning Classification, Medical	Very time consuming (HOG), Time consuming(invasive approach)	Less time consuming(UFL-ELM), Less time consuming(Wavelet base K mean clustering analysis)
Accuracy	Network traffic Classification, Object recognition	Nearby 79% (Port based and payload based),Low	Nearby 88% (K-means and Expectation maximization), High
Network load and false positive rate	Networking	Increase burden and false positive rate	Decrease burden and false positive rate
Semantic gap	Image recognition	Fails to remove semantic gap(CBIR without unsupervised)	Remove semantic gap(CBIR with unsupervised)
Multimodality	Robotics	Fails to capture(Switching GPs)	Able to capture(Hybrid systems)
Visualization	Road traffic network	With a single or more attribute	In form of Graph pattern problem(with multiple pattern)
Performance	Graph matching, Wireless communication	Less(Parameter learning), Less	High(Unsupervised Parameter learning), High (unsupervised secure sensing algorithm)
Cost, types of movements	Motor Movement	Expensive, specific types(Handcraft feature based)	Less expensive, multiple types(Auto encoder reconstruction)
Detection of threats	Cloud services	Difficult to detect	Automatically(Wavelet transform with unsupervised)
Relevant feature and performance	Sequential Data	Fails to find relevant feature and less performance	Easily find relevant features and high performance.
categorized level	Action categorized	Difficult	Easily
Accessing of high resolution image	Satellites Imaging	Face problem for accessing	Easily access

CONCLUSION

This paper evaluates existing technique and latest unsupervised learning techniques in diverse areas of interest such as handcraft feature based technique for evaluating movements of motor is very expensive and evaluate specific type, auto encoder technique of unsupervised learning removes these crisis and improve efficiency. In the entire area of interest, unsupervised techniques are discovered which gives good results as compared to the other techniques. These technique leads to future investigation such for learning hierarchy feature of image with representation low to high level feature.

REFERENCES

[1] Dao Lam, D. W. (2014). Unsupervised Feature Learning Classification Using An Extreme Learning Machine. *Neural Networks (IJCNN), The 2013 International Joint Conference on.* Dallas, TX, USA: IEEE.

[2] E. Steffi, I. J. (2017, march 13). Fish Image Recognition Using Benefit Based Partial Classification. *International Journal of Advanced Research Trends in Engineering and Technology* .

[3] Fan Zhang, B. D. (2014). Saliency-Guided Unsupervised Feature Learning for Scene Classification. *IEEE Transactions on Geoscience and Remote Sensing* , 2175 - 2184.

[4] Juan Carlos Niebles, H. W.-F. (2008). Unsupervised Learning of Human Action Categories Using Spatial-Temporal Words. *International Journal of Computer Vision , 79* (3), 299–318.

[5] Landress, A. D. (2016). A hybrid approach to reducing the false positive rate in unsupervised machine learning intrusion detection. *SoutheastCon, 2016.* Norfolk, VA, USA: IEEE.

[6] Lee, G. (2017). *Unsupervised Learning for Nonlinear PieceWise Smooth Hybrid Systems.* Computer Science.

[7] Marius Leordeanu, R. S. (2012). Unsupervised Learning for Graph Matching. *International Journal of Computer Vision* .

[8] Parisa Naraei, M. K. (2017). A Hybrid Wavelet based K-Means Clustering Approach to Detect Intracranial Hypertension. *Humanitarian Technology Conference (IHTC), 2017 IEEE Canada International.* Toronto, ON, Canada: IEEE.

[9] Prayook Jatesiktat, W. T. (2016). Unsupervised Anomalous Movement Detection using Autoencoder Reconstruction Error. *Net Regional Conference for Computer and Information Engineering 2016.*

[10] Senyan Yang, J. G. (2017). Analysis of traffic state variation patterns for urban road network based on spectral clustering. *Advances in Mechanical Engineering* .

[11] Singh, H. (2015). Performance Analysis of Unsupervised Machine Learning Techniques for Network Traffic Classification. *Advanced Computing & Communication Technologies (ACCT), 2015 Fifth International Conference o.* Haryana, India: IEEE.

[12] T. Dharani, I. L. (2013). A survey on content based image retrieval. *Pattern Recognition, Informatics and Mobile Engineering (PRIME), 2013 International Conference on.* Salem, India : IEEE .

[13] Wangyan Feng, W. Y. (2017). Wavelet transform and unsupervised machine learning to detect insider threat on cloud file-sharing. *Intelligence and Security Informatics (ISI), 2017 IEEE International Conference on.* Beijing, China : IEEE .

[14] Yang Li, Q. P. (2016). Achieving secure spectrum sensing in presence of malicious attacks utilizing unsupervised machine learning. *Military Communications Conference, MILCOM 2016 - 2016 IEEE.* Baltimore, MD, USA : IEEE.

[15] Yuhui Zheng, B. J. (2017). Student's t-Hidden Markov Model for Unsupervised Learning Using Localized Feature Selection. *IEEE Transactions on Circuits and Systems for Video Technology* , 1-1.