Machine Learning Recent Researches and its Challenges and Opportunities: A Review

Prasoon Purwar, Neha Bhardwaj
MTech Student, Assistant Professor
Department of CSE/IT,
Madhav Institute of Technology and Science, Gwalior, India

Abstract: We are in digital era where data comes in abundance, using self-learning algorithms from machine learning, we can turn available big data into knowledge. Machine learning is important aspect of data science that allows the systems to solve problems without being explicitly programmed. This brief review of machine learning covers basic of this technology and the recent researches have been done. This paper also describes the challenges faced and the future opportunities of machine learning. Some technical challenges are like, creating algorithms that can computationally scale to large data sets, designing machine learning algorithm that can learned and gives better results without requiring large volume of labeled data and to improving the hardware to implement big machine learning systems.

Keywords: Machine Learning, Deep Learning, Classification, Computer Vision.

1. Introduction

In today's digital era, with the emergence of technology, the digital data across all sectors growing exponentially. As volumes of data increases, new methods to extract meaningful knowledge from the data are emerging. The last few years have seen exciting advances in machine learning, which have inflated the potentiality across a suite of applications. Increased data availability has allowed machine learning to be trained using large amount of examples, while increasing processing power has supported analytical potentiality of these systems. The large availability of data brings many opportunities and challenges to use of this data in various research applications. Machine learning also got greater power from the algorithmic advances which have been carried out in recent years. As a result of these technical developments in the field, increased availability of data and increased computing power, systems which only a few years ago struggled to achieve accurate results can now outperform humans at specific tasks. Different machine learning based systems have been emerged in recent years like image recognition system popularly used in social media, voice recognition systems, used by virtual personal assistant and recommendation system, used by ecommerce industry. Healthcare sector has widespread impact of machine learning. Systems developed using machine learning can be used for plethora of health related issues, like helping doctors to give effective diagnosis and treatment for certain medical conditions [Rohan Bhardwaj et al. 2017].

2. Machine Learning

Machine Learning is a particular method of data analytics that automate model building, as it relates to the development of models. Machine learning is the technology that allows systems to learn directly from examples, data and experiences. Machine learning systems are set a task and given a big volume of data for training and learning of how the given task can be achieved or from which to detect patterns. The system then trains itself and learns how to achieve best desired output. In some particular fields, machine learning is already able to achieve a higher level of performance than human. For example region based image classification model achieved 90% accuracy which is higher than any manual classification [M. S. Pawar et al. 2017].

Machine learning lies at the intersection of computer science, statistics and data science [Kajaree Das, Rabi Narayan Behera 2017]. It uses elements of each of these fields to process data in a way that can detect and learn from patterns, predict future activity, or make decision.

2.1. Branches of Machine Learning

There are three main approaches widely used in machine learning.
Supervised learning: In supervised learning, system is trained with example data which contains correct output or labels of the task. The labels classify each data point into one or more groups. The system learns how the training data is structured and uses this to predict the class of new unlabelled data. This type of supervise learning is classification model. Classification employs logistic regression, classification trees, support vector machines, artificial neural networks, random forest or other algorithms. Regression is supervised approach in which a learning model maps the data into real valued variable. The second type of machine learning is unsupervised learning.

Unsupervised learning: Unlike supervised learning, the training data does not contains the labels or targets in unsupervised learning. This type of learning explores data and attempt to develop some sort of patterns without any knowledge of predefined pattern. Unsupervised learning includes, clustering algorithms and dimensionality reduction algorithms. Clustering takes dataset covering various dimensions and partitions it into clusters according to some criteria or groups them into clusters having similar properties. A mostly used clustering algorithm is k-means clustering, which aims to make clusters of datasets so that each observation lies closest to mean of its cluster. Dimensionality reduction involves cutting away features that are redundant or irrelevant. It includes principal component analysis, linear discriminant analysis and multidimensional scaling. The third type of learning is reinforcement learning.

Reinforcement Learning: Reinforcement Learning focuses on learning from experience in order to maximize a reward and lies between supervised and unsupervised learning. In this type of learning, examples with optimal output are not given to the algorithm, but it must discover them by a process of hit and trial [Bishop CM. 2006].

Deep learning: In recent years deep learning has set an exciting new trend in machine learning. Deep learning, sub-category of machine learning inspired by artificial neural network. It is feature learning method that directly processes raw data extracts the features and learns from those features, which can be applied to target classification. Recently, deep learning made successful advances in machine learning and they have been applied to medical imaging [Litjens, G.J et al. 2017], computer vision [F. Zhang et al. 2018] and many other fields. Various deep learning architecture have been seen in literature includes recurrent neural network (RNNs), deep belief network (DBNs), convolution neural network (CNNs), deep autoencoders [Daniele Ravı et al. 2017].

3. Literature Review

In past few years many papers have been published and accepted in the field of machine learning and its sub branch deep learning. As the use of internet increasing for shopping, watching movies or listening music, many websites are trying to give their user recommendations for their liking using recommendation systems. Nawrocka A. et al, (2018) presented utilization of recommendation system based on collaborative filtering algorithms. The presented algorithm was based on similarity of objects and users to render recommendations. Algorithm evaluated on MovieLens dataset gives less root mean square error.

Machine learning algorithms are applied widely in text mining research to process and extract meaningful knowledge or information from unstructured text data which consist of big documents from several sources like news article, research papers, web pages and more. Wang, Z., & Qu, Z. (2017), proposed model for web text classification in web information retrieval by merging traditional machine learning algorithm i.e. support vector machine with deep learning architecture i.e. convolutional neural network. Noviantho et al. (2017), built a classification model for observing cyberbullying, since the detection of these type of attacks in online platform becomes very important. Authors constructed a model to classify cyberbullying on text messages conversations using text mining methods. Social media has been used for thoughts sharing with others however this platform has been misused by some malicious users. Sharmin et al. (2017), proposed model for spam detection in social media by using machine learning methods. This model categorizes the text into two classes viz spam and not-spam. For this research Youtube comment data was used to filter out spam comments, the ensemble classifier performs better in most of the cases.

Image analysis and computer vision have been important in scientific and industrial applications. Several advancements have been made in computer vision and image analysis. Kajale, Renuka et al. (2017), presented a model for accurate and efficient recognition system for hand written characters and printed nameplates and Zhang et al. (2018), recognition system for accurate automatic detection of characters printed on machines electrical box based on support vector machine. Recognition rate obtained from SVM based multiclass classifier compared with template matching method, SVM based model performs better over template matching, since template matching method only performs parallel movement. Object recognition is also receiving close review in research, Hayat S. et al. (2018), proposed a deep neural network based multi class object recognition system which categorize images of object to nine different object classes. Model was based on convolutional neural network for the recognition process, the CNN was fine-tuned using back propagation algorithm, model evaluated in terms of accuracy and achieves better performance than compared methods. Computer vision based phenotyping gives the ability to analyse the quantitative plant physiology, using machine learning based approach, Islam, M. et al. (2017) combines machine learning and image processing to present a model to diagnose potato plant leaf diseases using SVM. Model first finds the region of interest that contains visible diseases symptoms and segments these region of interest to train using multiclass SVM. In CNN based computer vision machine learning algorithms require feature selection and extraction plays a vital role for several tasks. Feature selection can optimize the efficiency of algorithm and can reduce the complexity of algorithm. F. Zhang et al. (2018), proposed a model for data driven convolutional feature selection for machine learning algorithm. This scheme selects the features related to object and the other features are discarded which is driven by data by assigning an adaptive weight to each feature map.

With the success in computer vision, deep learning application in clinical imaging data offers broad review of research. Wang, et al. (2017), presented a chest x-ray data sets which comprises 10,000 frontal view X-ray images, with label of eight disease and also demonstrated CNN based network to automatically locate and detect the thoracic disease in the image. With some improvement in chest-x-ray data set and model, Rajpurkar, Pranav et al. (2017), presented cheXNet a CNN based model to detect pneumonia in chest x-ray images. In similar approach, Rajpurkar, Pranav et al (2017), introduced MURA, a large dataset of musculoskeletal radiographs from different studies and developed a 169- layer CNN model which was trained with the introduced datasets to detect and localize abnormalities in the radiograph images. Deep learning also used to segment of dermoscopic skin lesions. Venkatesh, et al. (2018), proposed model for an automatic skin lesion segmentation based on deep learning architecture and Patiño, D., et al. (2018), also proposed a model for melanoma detection in skin lesion image data set using simple linear iterative clustering (SLIC) algorithm to segment the skin lesion on dermoscopic images. Since skin cancer is most dangerous disease in many part of world, detection of this type of disease in prior stages is more important, former model uses UNet with residual connection for segmentation of lesions region, classification of each pixel based on pertinent geometrical and color features optimized using Ant Colony Optimization. Ke Yan, et al (2018), proposed to mine and harvest abundant retrospective medical data to build a large scale lesion image dataset and developed DeepLesion, a dataset with 32,735 lesions in 32,120 CT slices from 10,594 studies of 4,427 unique patients. There are variety of lesion types in this dataset, such as lung nodules, liver tumors, enlarged lymph nodes etc. Using DeepLesion, they have trained a universal lesion detector that can find all types of lesions with one unified framework.

Recently, machine learning also has been applied to process electronic health records, including both structured like diagnosis, medications, laboratory test and unstructured like clinical notes. The big part of this literature applied to process EHR's of health care systems using conventional machine learning or by deep learning. Many works applied machine learning algorithms to predict disease from patient records. T. Christensen, et al. (2018), proposed disease prediction model using code embedding learning, Code2Vec for learning. Then, this learning is used with the XGBoost to predict the disease from health records. Since, diagnosis recommendation plays important role, but supervised algorithms require many labeled instances of data which is not be generated or available. Ahmed, Ishtiaq et al. (2018), used the semi-supervised method for recommendation of disease diagnosis. Proposed model uses semi-supervised method for recommending labels of diagnosis with the help of clustering and frequent pattern mining. The whole model is implemented on DATAVIEW. Differently, Brisimi, Theodora S. et al. (2018), developed a predictive analytics to predict the hospitalization due to two severe chronic diseases i.e. heart disease and diabetes. Authors evaluated some baseline methods including SVM, random forest. They developed alternating clustering and classification approach to predict hospitalization due to chronic diseases which gives promising results compared to other approaches. Apart from these methods for prediction and recommendation, generation of time series medical data which contains required parameters from patient care systems is also important, which can be used for many research applications. Shamsuddin, R, et al. (2018), proposed virtual patient model (VPM) for generating time series healthcare data. VPM depicts the technique which synthesizes time series data to virtual patient data. Machine learning algorithms are widely used in bioinformatics for classification of different diseases. Sharma A. and R. Rani (2017), proposed a scheme for classification of cancer cell lines based on their genetic similarity and the type of cancer. Authors used hybrid algorithms which includes neural network and support vector machine and model gives good result for classification of cancer types. In other research in microarray breast cancer classification, S. Turgut, et al. (2018), used several classification algorithms for classification of microarray breast cancer and compared each algorithm with other algorithms and support vector machine gives the better results for the classification of breast cancer on two different datasets.

Table 1, summarizes the papers referred in literature review, in general highlighting the type of machine learning algorithm used and the data considered. We have also considered some papers based on deep learning used in computer vision or in clinical image analysis in this review, which also summarized in this table.

Table 1, Summary of literature review

References	*Application*	*Model*	*Data Set*
Nawrocka A. et al, 2018	Recommendation System	Collaborative Filtering	MovieLens
Wang, Z., & Qu, Z. 2017	Web Text Classification	CNN-SVM	20-NewsGroup corpus
Noviantho et al. 2017	Detection of cyberbullying in online platforms	SVM with poly kernel	Formspring
Sharmin, S., & Zaman, Z. (2017)	Detection of spam messages in social media platforms	Ensemble Classifier	Collected from UCI data repository
Kajale, Renuka et al. 2017	Character Recognition of handwritten and printed name plates	Decision Tree	Character Dataset
Zhang et al. 2018	Character Recognition for automotive electrical box component based on computer vision	SVM	Electrical characters
Hayat S. et al. 2018	Multiclass object recognition model	CNN	Caltech-101
Islam, M. et al. 2017	Potato Plant Leaf Disease Detection	SVM with RGB imaging	Plant Village
F. Zhang et al. 2018	Data Driven Feature Selection for Machine Learning Algorithms	Based on SRDCF	OTB-2013 and OTB-2015
Rajpurkar, Pranav et al. 2017	Radiologist-Level Pneumonia Detection on Chest X-Rays	CNN	ChestX-ray14
Rajpurkar, Pranav et al 2017	Abnormality detection in musculoskeletal radiographs	CNN	MURA
Venkatesh, et al. 2018	Deep Residual Architecture for Skin Lesion Segmentation	U-net and residual network	Skin Lesion- ISIC 2017
Patiño, D., et al. 2018	Melanoma Detection in Skin Lesion	SLIC	Skin Lesion- ISIC 2017
Ke Yan, et al 2018	Automated Mining Of Large Scale	VGG-16	DeepLesion

	Lesion Annotations And Universal Lesion Detection		
T. Christensen, et al. 2018	Machine Learning Methods for Disease Prediction	XGBoost	Electronic Health Record
Ahmed, Ishtiaq et al. 2018	Diagnosis Recommendation Using Semi-supervised approach	Clustering and Frequent pattern mining	Electronic Health Record
Brisimi, Theodora S. et al. 2018	Predicting Chronic Disease Hospitalizations from Electronic Health Records	Alternating clustering and classification (ACC)	Electronic Health Record
Sharma A. and R. Rani 2017	Classification of cancerous profile	NN- SVM	Cancer X-gene
Turgut S., et al. 2018	Microarray Breast Cancer Data Classification	Classification algorithms	Microarray Breast Cancer

CNN- convolutional neural network, SVM- support vector machine, NN- neural network, SRDCF- spatially regularized discriminative correlation filters, SLIC- simple linear iterative clustering

4. Analysis

The previous section reviews some recent literatures of machine learning. Above table describes the model presented by authors for different applications using many machine learning algorithms. The model of Wang, Z., & Qu, Z. (2017) for web text classification acquires the 92.5% accuracy for combined CNN and SVM than applying CNN or SVM separately on data set. The cyberbullying text classification model of Noviantho et al. (2017) using SVM-Poly algorithm achieves 99.41% accuracy for 2 classes classification and 97.81% for 4 classes. Character recognition of numbers written on electrical box, recognition rate of SVM is 99.40% which is better than compared to template matching model i.e. 98.3% [Zhang, L. et al. 2018]. CNN architecture used by many researchers for their research gives better performance in different application. CNN gives 90.12% result than classical BOW methods for multiclass object recognition [Hayat, S. et al. 2018]. In other model for pneumonia detection on Chest X-Rays the CNN based model achieves an F1 score of 0.435 (95% CI 0.387, 0.481) which is better than radiologist average analysis [Rajpurkar, Pranav et al. 2017]. Similarly for musculoskeletal radiograph studies of the upper extremity gives good performance than manual radiologist evaluation [Rajpurkar, Pranav et al. 2018]. DeepLesion dataset of radiological images used for universal lesion detection by implementing VGG-16 architecture achieves a sensitivity of 81.1% with five false positive per image [Ke Yan et al. 2018]. DeepLesion is a large dataset of radiological image can be used for different medical lesion researches for developing and training of deep models for lesion detection. Skin lesion segmentation for detection of melanoma for treating skin cancer is also a good research area. The deep model proposed using UNet with residual connection achieves accuracy of 0.936 and sensitivity 0.83 [Venkatesh, G. M et al. 2018]. Electronic health records are the broad area of research, the classification model for prediction of hospitalizations due to chronic diseases based on alternating clustering and classification (ACC) achieves approx. 77% on heart disease data and approx. 76% on diabetes data.

5. Challenges and Opportunities

Machine Learning is a sparkling field of research with a wide range of fields for further development across several methods and applications areas. Recent advances in the area have created new research challenges, which include both technical advances in the field and the societal challenges related to machine learning. Some technical challenges are like, creating algorithms that can computationally scale to large data sets, designing machine learning algorithm that can learned and gives better results without requiring large volume of labeled data and to improving the hardware to implement big machine learning systems. In many applications like safety or critical applications, the correctness of predictions made by machine learning systems needs to be verifiable to best known level. Data pre processing for applying machine learning algorithm also presents an important issue. Machine Learning systems need to deal with realties of real world data. Data sets can also have missing values and outliers, data comes from various sources and different formats, suffers from several form data corruption. Most of the time was spent in cleaning of data. To overcome this we need to develop efficient algorithms for data pre-processing or for gathering and generation of data. Develop standards for the processing and sharing of data, which allow quality of data to be used. Privacy is also an important issue in machine learning systems since they use large amount of data to make predictions, inference and decisions. In several contexts, like data on individuals, it will be crucial to give respect to privacy of data. The issue relate to both the crude data and the data inferred by the systems. The solution to this problem may be homomorphic encryption of data through which all the processing to be done on completely encrypted data, so that crude data cannot be exposed to machine learning systems.

6. References

[1] Ahmed, Ishtiaq et al. "Diagnosis Recommendation Using Machine Learning Scientific Workflows." 2018 IEEE International Congress on Big Data (BigData Congress) (2018): 82-90.

[2] Bishop CM. Pattern recognition and machine learning. New York: Springer; 2006.

[3] Brisimi, Theodora S. et al. "Predicting Chronic Disease Hospitalizations from Electronic Health Records: An Interpretable Classification Approach." Proceedings of the IEEE 106 (2018): 690-707.

[4] Christensen T., A. Frandsen, S. Glazier, J. Humpherys and D. Kartchner, "Machine Learning Methods for Disease Prediction with Claims Data," 2018 IEEE International Conference on Healthcare Informatics (ICHI), New York City, NY, USA, 2018, pp. 467-4674.

[5] Daniele Ravı, Charence Wong, Fani Deligianni, Melissa Berthelot, Javier Andreu-Perez, Benny Lo, Guang-Zhong Yang, "Deep Learning for Health Informatics", IEEE JOURNAL (2017).

[6] F. Zhang, W. Li, Y. Zhang and Z. Feng, "Data Driven Feature Selection for Machine Learning Algorithms in Computer Vision," in IEEE Internet of Things Journal (2018).

[7] Hayat, S., Kun, S., Tengtao, Z., Yu, Y., Tu, T., & Du, Y. (2018). A Deep Learning Framework Using Convolutional Neural Network for Multi-Class Object Recognition. 2018 IEEE 3rd International Conference on Image, Vision and Computing (ICIVC).

[8] Islam, M., Anh Dinh, Wahid, K., & Bhowmik, P. (2017). Detection of potato diseases using image segmentation and multiclass support vector machine. 2017 IEEE 30th Canadian Conference on Electrical and Computer Engineering (CCECE).

[9] Kajaree Das, Rabi Narayan Behera, "A Survey on Machine Learning: Concept, Algorithms and Applications," International Journal of Innovative Research in Computer and Communication Engineering, Vol. 5, Issue 2, February 2017.

[10] Kajale, Renuka et al. "Supervised machine learning in intelligent character recognition of handwritten and printed nameplate." 2017 International Conference on Advances in Computing, Communication and Control (ICAC3) (2017): 1-5.

[11] Ke Yan, Xiaosong Wang, Le Lu, Ronald M. Summers, "DeepLesion: automated mining of large-scale lesion annotations and universal lesion detection with deep learning," J. Med. Imag. 5(3), 036501 (2018), doi: 10.1117/1.JMI.5.3.036501.

[12] Litjens, G.J., Kooi, T., Bejnordi, B.E., Setio, A.A., Ciompi, F., Ghafoorian, M., Laak, J.V., Ginneken, B.V., & Sánchez, C.I. (2017). A survey on deep learning in medical image analysis. Medical image analysis, 42, 60-88 .

[13] M. S. Pawar, L. Perianayagam and N. S. Rani, "Region based image classification using watershed transform techniques," 2017 International Conference on Intelligent Computing and Control (I2C2), Coimbatore, 2017, pp. 1-5.

[14] Nawrocka A., A. Kot and M. Nawrocki, "Application of machine learning in recommendation systems," 2018 19th International Carpathian Control Conference (ICCC), Szilvasvarad, 2018, pp. 328-331. doi: 10.1109/CarpathianCC.2018.8399650

[15] Noviantho, Isa, S. M., & Ashianti, L. (2017). Cyberbullying classification using text mining. 2017 1st International Conference on Informatics and Computational Sciences (ICICoS). doi:10.1109/icicos.2017.8276369.

[16] Patiño, D., Avendaño, J., & Branch, J. W. (2018). Automatic Skin Lesion Segmentation on Dermoscopic Images by the Means of Superpixel Merging. Lecture Notes in Computer Science, 728–736. doi:10.1007/978-3-030-00937-3_83.

[17] Rajpurkar, Pranav et al. "CheXNet: Radiologist-Level Pneumonia Detection on Chest XRays with Deep Learning." CoRR abs/1711.05225 (2017): n. pag.

[18] Rajpurkar, Pranav et al. "MURA: Large Dataset for Abnormality Detection in Musculoskeletal Radiographs." (2018), arXiv:1712.06957.

[19] Rohan Bhardwaj, Ankita R. Nambiar, Debojyoti Dutta "A Study of Machine Learning in Healthcare" 2017

IEEE 41st Annual Computer Software and Applications Conference.

[20] Shamsuddin, R., Maweu, B.M., Li, M., & Prabhakaran, B. (2018). Virtual Patient Model: An Approach for Generating Synthetic Healthcare Time Series Data. 2018 IEEE International Conference on Healthcare Informatics (ICHI), 208-218.

[21] Sharma A. and R. Rani, "Classification of Cancerous Profiles Using Machine Learning," 2017 International Conference on Machine learning and Data Science(MLDS), Noida, India, 2018, pp. 31-36.

[22] Sharmin, S., & Zaman, Z. (2017). Spam Detection in Social Media Employing Machine Learning Tool for Text Mining. 2017 13th International Conference on Signal-Image Technology & Internet-Based Systems (SITIS), 137-142.

[23] Turgut S., M. Dağtekin and T. Ensari, "Microarray breast cancer data classification using machine learning methods," 2018 Electric Electronics, Computer Science, Biomedical Engineerings' Meeting (EBBT), Istanbul, 2018, pp. 1-3.

[24] Venkatesh, G. M., Naresh, Y. G., Little, S., & O'Connor, N. E. (2018). A Deep Residual Architecture for Skin Lesion Segmentation. OR 2.0 Context-Aware Operating Theaters, Computer Assisted Robotic Endoscopy, Clinical Image-Based Procedures, and Skin Image Analysis, 277–284.

[25] Wang, Z., & Qu, Z. (2017). Research on Web text classification algorithm based on improved CNN and SVM. 2017 IEEE 17th International Conference on communication Technology (ICCT). doi:10.1109/icct.2017.8359971.

[26] Wang, Xiaosong, Peng, Yifan, Lu, Le, Lu, Zhiyong, Bagheri, Mohammadhadi, and Summers, Ronald M. Chestx-ray8: Hospital-scale chest x-ray database and benchmarks on weakly-supervised classification and localization of common thorax diseases. arXiv preprint arXiv:1705.02315, 2017.

[27] Zhang, L., Pang, D., & Ma, P. (2018). Character recognition for automotive electrical box components based on Machine vision. 2018 International Conference on Advanced Mechatronic Systems (ICAMechS). doi:10.1109/icamechs.2018.8506926.