International Journal of Research in Social Sciences
Vol. 8 Issue 2, February 2018,
ISSN: 2249-2496 Impact Factor: 7.081
Journal Homepage: http://www.ijmra.us, Email: editorijmie@gmail.com
Double-Blind Peer Reviewed Refereed Open Access International Journal - Included in the International Serial Directories Indexed & Listed at: Ulrich's Periodicals Directory ©, U.S.A., Open J-Gate as well as in Cabell's Directories of Publishing Opportunities, U.S.A

Machine learning algorithms: present supervised and unsupervised learning techniques

Sweety Kataria

Assistant Professor, Kalindi College

ABSTRACT

Machine Learning is learning from data. Machine learning is to improve computers automatically through previous learning and ML is an growing field at the intersection of computer engineering and methods of statistics. ML methods are mainly classified as two main groups i.e. Supervised and Unsupervised learning. The algorithms of supervised learning and widely used classification techniques areSVM, Decision tree, Random forest, k-Nearest Neighbors, Neural networks, Polynomial regression. Regression method and Classification Techniquesare most important methods in supervised ML. Commonly used regression ML methods are: Linear, Logistic, Polynomial, Lassoand Multivariate Regression methods. Different classification techniques widely used are discriminant analysis, Neural Networksand SVM. All clustering algorithms are unsupervised learning techniques i.e. K – Means clustering, Hierarchical clustering like agglomerative hierarchical clustering, Density Based Clustering, used to identify distinctive clusters in the data.ML techniques have been used widely in various applications andareas.It has become an extremely popular tool for material science, medical sciences for diseases like cancer, health system and in computer sciences like in Defect prediction in software development etc.

Keywords: Deep Learning, Machine Learning, Defect prediction, Unsupervised learningandSupervisedlearning.

INTRODUCTION

Machine Learning is an artificial intelligence technique, mainly relates to learning from data. Machine learning is used for patterns or to extract important information from the data (Richert, 2013). Machine learning is to improve computers and programme to work automatically through earlier experience. It is most increasing technical fields at the intersection of computer engineering and methods of statistics, at the core of artificial intelligence and data science (Jordan and Mitchell , 2015). The main purpose of machine learning is to learn from the data. Industries of diverse arear from medicine to military are using machine learning to interpret the relevant information from the data. MLalgorithms has potential in solving range of engineering problems and are useful where problem domains are not well defined, to develop efficient algorithms.

Machine learning (ML)methodsmainly grouped into two main categories i.e Unsupervised learning andSupervisedlearning methods. Supervised learning relates to learning from earlier data with an known outcome, while unsupervised learning is learning from data without outcome. In Supervised learning input and output variable, both are required and algorithm to learn the features from the input data to generate known outputdata. In Unsupervised learning only input data and no output mean unsupervised learning learn from data without known outcome. So the supervised techniques cannot be effectively used for efficient models, if previous data is limited. (Fig.1)

There has been a large variety of ML algorithms or classifiers for supervised learning from variety of algorithm families i.e. decision trees, classification rules, neural networks and probabilistic classifiers. ML algorithms have wide range of applications and have significance role in solving wide range of engineering problems i.e. prediction of failure, error and defect. ML includes different learning methods/ techniques such as MLs artificial neural network, concept learning, Bayesian network, reinforcement learning, instance-based learning,genetic algorithms & programming, , decision trees, inductive logic programming, and analytical learning, many of these are demonstrated in Fig. 2. Studies on many ML techniques by many researcher have been reported (Xi Tan et al., 2011; Kriti Purswani et al., 2013; Gabriela Czibula et al., 2014) and showed that in ML techniques such as unsupervised learning methods and supervised learning methods have been used for building software defect prediction models. Supervised learning approacheshave potential and give more accurate and efficient for software defects predicting approach, if significant data available. The Supervised learning techniques cannot be effectively used for efficient models, if learning data is inadequate. Mostly effective and economical method for prediction defects in software development are learning from prior mistakes and prevent in further processes.ML techniqueshave significant role where problemsare not well defined and human mind is limited. ML based models uses data mining methods &algorithm and statistical methods for defect classification.

In the present review, the main objective is to present review of the literature pertains to evaluation of machine learning techniques software defect prediction models.

Fig. 1: Training and Testing phases of any sample image.

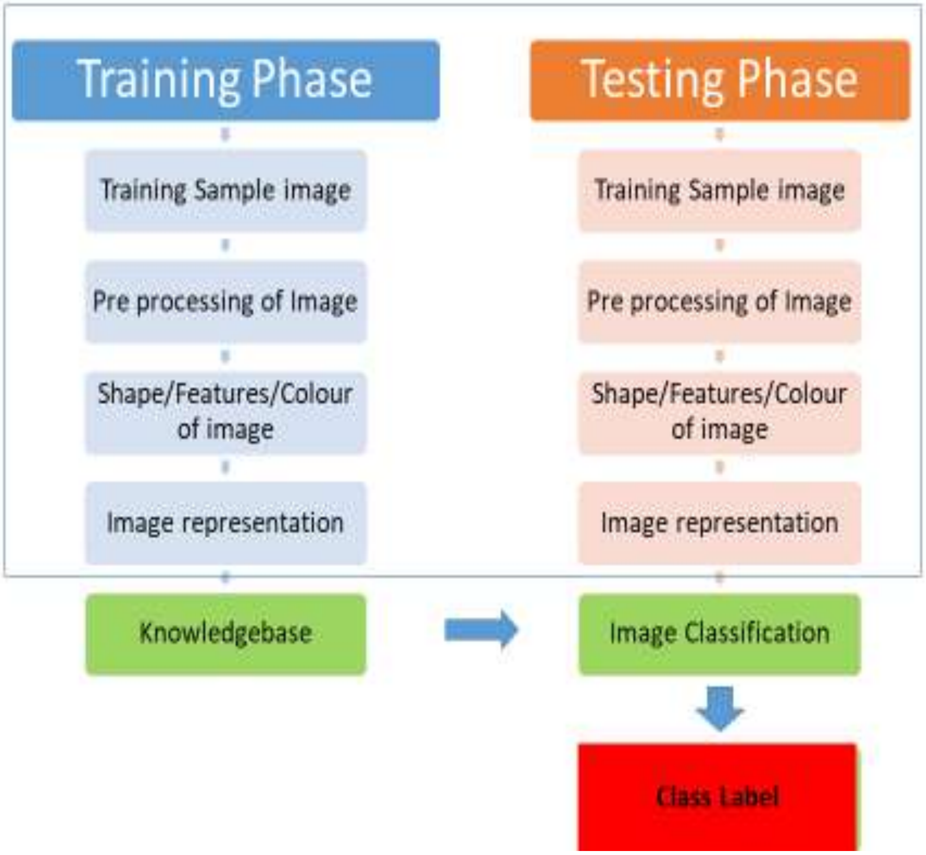

Table 1: Steps for class label of image in Learning.

A. Training Phase
• Training Sample image • Pre processing of Image • Shape/Features/Colour of image • Image representation

B. Testing Phase

- Training Sample image
- Pre processing of Image
- Shape/Features/Colour of image
- Image representation

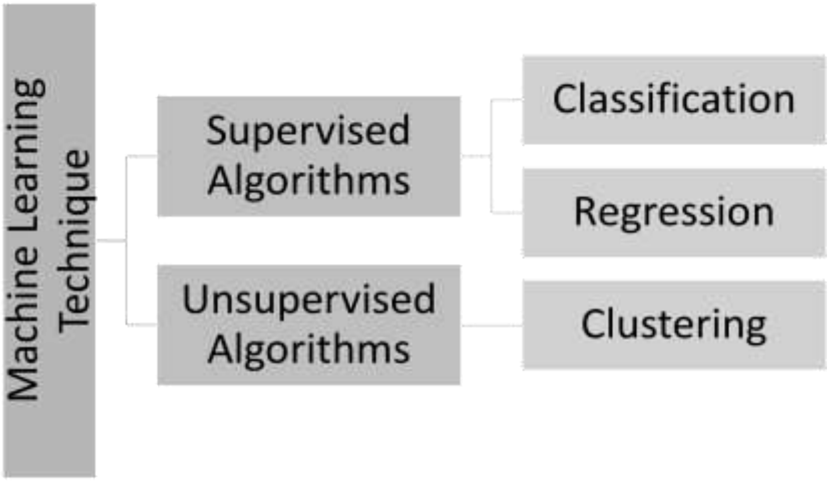

Fig 2: Machine learning algorithms

Machine Learning (ML):

ML is enabling computer programs automaticto improve performance at through experience or training. There are several studies by researchers using ML in many areas.

Supervised Learning:

The Supervised learning is learning from previous data/examples with defined output/outcome. Supervised learning is a tool for the classify and processing of the data using machine. Supervised learning methods needs data set which has been classified, for learning algorithm. ML is already trained for input termed as supervised learning, that will be target for new input after training. Targets expressed in some classes are called classification methods.These ML algorithmsuse data as the basis for predicting the classification of other unclassified data termed or unlabeled data.

The classification and regression algorithms of ML techniques are supervised learning. Widely reported supervised learning algorithms are Decision tree classification algorithm,to measure the quality of a split; Support vector machine, machine with associated learning algorithms (SVM), k-Nearest Neighbors(KNN), to solve both classification and regression problems, Naive Bayes,based on Bayes' Theorem; Random forest,(random decision forests): ensemble learning method for classification, regression Neural networks,can recognize relation between large data; Polynomial regression, relationship between a dependentand independent variableas nth degree polynomial.. (Kramer 2013; Anderson, 1995; Hastie et al, 2009; Caruana and Niculescu-Mizil, 2006)

Table 2: The important algorithms of Supervised learning includes:
• Classification algorithm ; Decision tree, to measure the quality of a split
• Support vector machine (SVM): with associated learning algorithms
• k-Nearest Neighbors (KNN): to solve both classification and regression problems.
• Naive Bayes algorithms: based on Bayes' Theorem
• Random forest (random decision forests): ensemble learning method for classification, regression
• Neural networks : can recognize relationbetween large data
• Polynomial regression: relationship between a dependentand independent variableas nth degree polynomial.
• SVM for regression

Most important and widely used supervised learningtechniques are Regressionand Classification Techniques.

(a) Regression method:

A regression techniquesare predictive ML technique which helps in the prediction of association of dependents i.e. target and independent variables. Commonly used regressiontechniques are:

- Linear Regression
- Logistic Regression
- Polynomial Regression
- Lasso Regression
- Multivariate Regression

The Linear regression is a category of techniques under supervised learning, mainly used for prediction, for forecastingand for finding relationships in data sets. This learning technique is widely used supervised learning technique. For example, this technique is use to determine a relationship between a particular drug and tumours. (Chou 2012)

(b) Classification Techniques:

The classification techniques are mainly for predicting a qualitative response and forrecognition of pattern in data. Different classification techniques or classifiers (Liao and Carin, 2005, Phyu, 2009; Xanthopoulos et al, 2013; Mitchell, 2014) are as below:

- Logistic regression,
- Linear discriminant analysis,
- K-nearest neighbors,
- Trees,
- Neural Networks, and
- Support Vector Machines.

Unsupervised learning:

In the group of unsupervised learning technique, no previous data set is trained and is only trained with a input data set termed unsupervised learning [Chung et al., 2013], provides structure or relationships betweeninputs. Unsupervised learning uses unlabelled and unclassified, and categorized training data. The unsupervised learning is for discover hidden and interesting patterns in unlabeled data.

(a) Clustering:

The Clustering techniques are mostly used, common and important unsupervised learning algorithm, are used to explore and analyse data to find out patterns or groupings in the data which shares common characters. (Nunez-Iglesias, 2013). Unsupervised models include clustering techniques, use different strategies for dividing data into groups or clusters. (McCue, 2014; Nunez-Iglesias, 2013)

Clusteringalgorithms: Clustering unsupervised learning techniques.

- K – Means clustering
- Hierarchical clustering
- Make Density Based Clustering.

The main algorithms for clustering are as below: are also termed as dimensionality reduction techniques

K-Means Clustering:

Clustering is a type of unsupervised learning technique, it creates groups automatically. The items which possesses similar features, values or character are grouped in same cluster. This algorithm is called k-means because it creates k distinct clusters. (Dey, 2016) The mean of the values in a particular cluster is the center of that cluster [Shalev-Shwartz et al., 2011].

Principal Component Analysis:

Principal Component Analysis (PCA) is used to reduction in the dimension of the data to make the analysis faster and easier. PCA can reduce the data is being plot in a graph into two axes. So PCA can be applied to the data to convert into 1D (Dey, 2016).

Table 3: Different Machine learning algorithms:

Regression: -	***Classification: -***	***Clustering: -***
• Linear Regression • Non – linear Regression • Logistic Regression • Multivariate Regression	• Naïve Bayes • Support Vector Machine • Random Forest • Decision Trees • Neural Networks • Nearest Neighbour	• K-Means Clustering • Hierarchical Clustering • Probabilistic Clustering • Density – based Clustering

RELATED WORK

Ongoing through the reviewed literature, Machine learning algorithms have been found to important growing field used widely in various applications and areas (Freitag, 1998; Wang and Zhou , 2009). It has becomean extremely popular tool for material science (Yosipof et al., 2015), medical sciences for diseases like cancer (Erickson et al., 2017; Fatima, and Pasha, 2017) as well as in computer sciences like in Defect prediction etc. (Malhotra, 2016; Wahono, 2015; Shepperd et al., 2014; Bibi et al., 2006; Wang et al., 2010).

Bibi et al., 2006, studied Software Defect Prediction Using Regression via Classification.

Wahono et al., 2015, worked on systematic literature review of software defect prediction.

A framework for taking of machine learning in industry for software defect prediction and increasing use of Machine learning algorithms being used in a variety of application domains including software engineering are reviewed by Rana et al., 2014

Challagulla et al., 2008, review the software defect prediction techniques with special emphasis on machine learning based methods. These techniques are applied to three real-time defect data sets obtained from NASA's MDP (Metrics Data Program) data repository.

Wang et al., 2010, conducted a comparative study of ensemble feature selection techniques for software defect prediction.

Wang, and Zhou, 2009, reviewed techniques with application in different field of machine learning and gives details its application.

Chen and et al [2010] reviewed earlier studies in the field of defect management and software prediction and introduce a novel method for defect prediction using data mining techniques and claim that their proposed model is able to lead the developmental stages of a new software.

Gayathri and Sudha (2014) explored an enhanced multilayer perceptron neural network and performed by comparative analysis for software systems and then tested by NASA's Metrics Data Program.

Shepperd et. Al., 2014, conducted a meta-analysis of all relevant, high quality primary studies of defect prediction to determine what factors influence predictive performance.

Ozturk et al., 2015, studied clustering algorithms for defect prediction. Four variants of K-mean clustering algorithm were taken into consideration using four real life datasets. Authors claimed that K-mean++variant gives better results than other K-mean variants.

Kim and Kim (2015) studied recurrent neural network to intrusion detection with hessian free optimization. Gabriela Czibula et al., 2014 studied Software defect prediction using relational association rule mining and reported a novel classification model regarding relational association rules mining.David and Netanyahu (2015) worked deep learning for automatic malware signature generation and classification. Pascanu et al., (2015) studied Malware classification with recurrent networks. Yuan et al., (2016) worked on android malware characterization and detection using deep learning.

Dey, 2016 studied the multivariate technique i.e. Principal Component Analysis (PCA) application in ML. PCA is used for reduction in the dimension of the data, by grouping of variable in accordance to the characters to make the analysis faster and easier. PCA can reduce the data is being plot in a graph into two axes. So PCA can be applied to the data to convert into 1D

Malhotra, 2016, studied empirical framework for defect prediction using machine learning techniques with Android software.

Erickson et al., 2017 studied and reviewed Machine learning for medical imaging.

Erickson et al., 2017, worked on ML techniques in disease diagnosis and Survey of machine learning algorithms for disease diagnostic.

The reviewed literature showed suitability of ML supervised and unsupervised learning techniques for defect prediction areactive areas of research software development have potential application in all areas, in software as well as health system etc.

CONCLUSION

Machine Learning, is artificial intelligence, is to improve computers automatically. In the current scenario, it is an evergreen research field and is rapidly increasing fields at the intersection of computer engineering and statististical methods. Machine learning algorithms have been used widely in various applications and areas. It has become an extremely popular tool for most of filed including diseases like cancer and health system as well as in computer sciences like in Defect prediction in software development.On-going through the reviewed literature, it is observed that suitability of Machine Learning supervised and unsupervised learning methods for defect prediction is an active area of research.ML have importance in problems which are not well defined and human mind is limited. ML based algorithms uses data mining techniques and algorithm based statistical methods for software defect classification in development of software.

REFERENCES

Challagulla, V.U.B., Bastani, F.B., Yen, I.L. and Paul, R.A., (2008). Empirical assessment of machine learning based software defect prediction techniques. International Journal on Artificial Intelligence Tools, 17(02), 389-400.

Chen, Y., Shen, X.H., Du, P. and Ge, B., (2010), February. Research on software defect prediction based on data mining. In 2010 The 2nd International Conference on Computer and Automation Engineering (ICCAE) (Vol. 1, pp. 563-567). IEEE.

Chug, Anuradha, and Shafali Dhall. "Software defect prediction using supervised learning algorithm and unsupervised learning algorithm." (2013): 5-01.

Czibula, G., Marian, Z. and Czibula, I.G., (2014). Software defect prediction using relational association rule mining. Information Sciences, 264, 260-278.

David OE, Netanyahu NS (2015) Deepsign: deep learning for automatic malware signature generation and classification. In: 2015 international joint conference on neural networks (IJCNN). IEEE

Rana, R., Staron, M., Hansson, J., Nilsson, M. and Meding, W., (2014), August. A framework for adoption of machine learning in industry for software defect prediction. In 2014 9th International Conference on Software Engineering and Applications (ICSOFT-EA) (383-392). IEEE.

Dey, A., (2016). Machine learning algorithms: a review. International Journal of Computer Science and Information Technologies, 7(3), 1174-1179.

Gabriela Czibula, Zsuzsanna Marian, Istvan Gergely Czibula, "Software defect prediction using relational association rule mining", Information Sciences, Volume 264 p 260-278, April (2014).

Gayathri, M. and Sudha, A., (2014). Software defect prediction system using multilayer perceptron neural network with data mining. International Journal of Recent Technology and Engineering, 3(2), 54-59.

Ishitaki et al (2017) studied application of deep recurrent neural networks for prediction of user behavior in tor networks.

Ishitaki T et al (2017) Application of deep recurrent neural networks for prediction of user behavior in tor networks. In: 2017 31st international conference on advanced information networking and applications workshops (WAINA). IEEE

Jordan, M.I. and Mitchell, T.M., (2015). Machine learning: Trends, perspectives, and prospects. Science, 349(6245), 255-260.

Kim J, Kim H (2015) Applying recurrent neural network to intrusion detection with hessian free optimization. In International workshop on information security applications. Springer.

Kriti Purswani, Pankaj Dalal, Dr. Avinash Panwar, Kushagra Dashora, "Software Fault Prediction using Fuzzy C-Means Clustering and Feed Forward Neural Network" International Journal of Digital Application & Contemporary Research vol 2, (1), (2013).

McCue, C., (2014)(. Data mining and predictive analysis: Intelligence gathering and crime analysis. Butterworth-Heinemann.

Öztürk, M.M., Cavusoglu, U. and Zengin, A., 2015. A novel defect prediction method for web pages using k-means++. Expert Systems with Applications, 42(19)6496-6506.

Pascanu R et al (2015) Malware classification with recurrent networks. In: 2015 IEEE international conference on acoustics, speech and signal processing (ICASSP). IEEE

Shalev-Shwartz, S., Singer, Y., Srebro, N. and Cotter, A., (2011). Pegasos: Primal estimated sub-gradient solver for svm. Mathematical programming, 127(1), 3-30.

Shepperd, M., Bowes, D. and Hall, T., (2014). Researcher bias: The use of machine learning in software defect prediction. IEEE Transactions on Software Engineering, 40(6), 603-616.

Xi Tan, Xin Peng, Sen Pan,Wenyun Zhao, "Assessing software quality by program Clustering and Defect Prediction" 18th Working Conference on Reverse Engineering (2011).

Yuan Z, Lu Y, Xue Y (2016) Droiddetector: android malware characterization and detection using deep learning. Tsinghua Sci Technol 21(1):114–123

Kramer, O., (2013). K-nearest neighbors. In Dimensionality reduction with unsupervised nearest neighbors (13-23). Springer, Berlin, Heidelberg.

Caruana, R. and Niculescu-Mizil, A., (2006), June. An empirical comparison of supervised learning algorithms. In Proceedings of the 23rd international conference on Machine learning (161-168).

Hastie, T., Tibshirani, R. and Friedman, J., (2009). Overview of supervised learning. In The elements of statistical learning (pp. 9-41). Springer, New York, NY.

Anderson, J.A., (1995). An introduction to neural networks. MIT press.

Chou, J.S. and Tsai, C.F., (2012). Concrete compressive strength analysis using a combined classification and regression technique. Automation in Construction, 24, 52-60.

Phyu, T.N., (2009), March. Survey of classification techniques in data mining. In Proceedings of the international multiconference of engineers and computer scientists (V. 1 no. 5).

Liao, X., Xue, Y. and Carin, L., (2005), August. Logistic regression with an auxiliary data source. In Proceedings of the 22nd international conference on Machine learning (505-512).

Xanthopoulos, P., Pardalos, P.M. and Trafalis, T.B., 2013. Linear discriminant analysis. In Robust data mining (27-33). Springer, New York, NY.

Mitchell, J.B., (2014). Machine learning methods in chemoinformatics. Wiley Interdisciplinary Reviews: Computational Molecular Science, 4(5), 468-481.

Nunez-Iglesias, J., Kennedy, R., Parag, T., Shi, J. and Chklovskii, D.B., (2013). Machine learning of hierarchical clustering to segment 2D and 3D images. PloS one, 8(8),.

Freitag, D., (1998), July. Information extraction from HTML: Application of a general machine learning approach. In AAAI/IAAI (517-523).

Wang, H., Ma, C. and Zhou, L., (2009), December. A brief review of machine learning and its application. In 2009 international conference on information engineering and computer science (pp. 1-4). IEEE.

Yosipof, A., Nahum, O.E., et al, (2015). Data Mining and Machine Learning Tools for Combinatorial Material Science of All-Oxide Photovoltaic Cells. Molecular informatics, 34(6-7), 367-379.

Erickson, B.J., Korfiatis, P., Akkus, Z. and Kline, T.L., (2017). Machine learning for medical imaging. Radiographics, 37(2), 505-515.

Fatima, M. and Pasha, M., (2017). Survey of machine learning algorithms for disease diagnostic. Journal of Intelligent Learning Systems and Applications, 9(01),1.

Shepperd, M., Bowes, D. and Hall, T., (2014). Researcher bias: The use of machine learning in software defect prediction. IEEE Transactions on Software Engineering, 40(6), 603-616.

Malhotra, R., (2016). An empirical framework for defect prediction using machine learning techniques with Android software. Applied Soft Computing, 49, 1034-1050.

Wahono, R.S., (2015). A systematic literature review of software defect prediction. Journal of Software Engineering, 1(1), pp.1-16.

Bibi, S., Tsoumakas, G., Stamelos, I. and Vlahavas, I.P., (2006), March. Software Defect Prediction Using Regression via Classification. In AICCSA (330-336).

Wang, H., Khoshgoftaar, T.M. and Napolitano, A., (2010), December. A comparative study of ensemble feature selection techniques for software defect prediction. In 2010 Ninth International Conference on Machine Learning and Applications (135-140). IEEE.