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 are SVM, Decision tree, Random forest, k-Nearest Neighbors, Neural networks, Polynomial regression. Regression method and Classification Techniques are 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 Networks and 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 and areas. 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 learning and Supervised learning.
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. ML algorithms 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) methods mainly grouped into two main categories i.e Unsupervised learning and Supervised learning 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 output data. 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 approaches have 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 techniques have significant role where problems are
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.

<table>
<thead>
<tr>
<th>A. Training Phase</th>
</tr>
</thead>
<tbody>
<tr>
<td>• Training Sample image</td>
</tr>
<tr>
<td>• Pre processing of Image</td>
</tr>
<tr>
<td>• Shape/Features/Colour of image</td>
</tr>
<tr>
<td>• Image representation</td>
</tr>
</tbody>
</table>
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 automatic to 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 algorithms use 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>
<thead>
<tr>
<th>Table 2: The important algorithms of Supervised learning includes:</th>
</tr>
</thead>
<tbody>
<tr>
<td>• Classification algorithm ; Decision tree, to measure the quality of a split</td>
</tr>
<tr>
<td>• Support vector machine (SVM): with associated learning algorithms</td>
</tr>
<tr>
<td>• k-Nearest Neighbors (KNN): to solve both classification and regression problems.</td>
</tr>
<tr>
<td>• Naive Bayes algorithms: based on Bayes' Theorem</td>
</tr>
<tr>
<td>• Random forest (random decision forests): ensemble learning method for classification, regression</td>
</tr>
<tr>
<td>• Neural networks : can recognize relation between large data</td>
</tr>
<tr>
<td>• Polynomial regression: relationship between a dependent and independent variable as nth degree polynomial.</td>
</tr>
<tr>
<td>• SVM for regression</td>
</tr>
</tbody>
</table>

Most important and widely used supervised learning techniques are Regression and Classification Techniques.

(a) Regression method:

A regression techniques are predictive ML technique which helps in the prediction of association of dependents i.e. target and independent variables. Commonly used regression techniques 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 forecasting and 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 for recognition 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 between inputs. 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)

Clustering algorithms: 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:

<table>
<thead>
<tr>
<th>Regression: -</th>
<th>Classification: -</th>
<th>Clustering: -</th>
</tr>
</thead>
<tbody>
<tr>
<td>Linear Regression</td>
<td>Naïve Bayes</td>
<td>K-Means Clustering</td>
</tr>
<tr>
<td>Non – linear Regression</td>
<td>Support Vector Machine</td>
<td>Hierarchical Clustering</td>
</tr>
<tr>
<td>Logistic Regression</td>
<td>Random Forest</td>
<td>Probabilistic Clustering</td>
</tr>
<tr>
<td>Multivariate Regression</td>
<td>Decision Trees</td>
<td>Density – based Clustering</td>
</tr>
<tr>
<td></td>
<td>Neural Networks</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Nearest Neighbour</td>
<td></td>
</tr>
</tbody>
</table>

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 become an 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.


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 plotted 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 are active 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 statistical 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


