A Smart Survey on Various Machine Learning Algorithms and It's Real Time Applications

Dr. Dushyantsinh B. Rathod[1], Dr. Ramesh Prajapati[2]

[1]Associate Professor & HOD, Computer Enginering,D.A.Degree Engineering and Technology,Mahemdabad,India

[2]Assistant Professor & HOD, Department of Computer Enginering,Indrashil University,Kadi,Gujarat,India

Abstract - *In today's world, most of the data is generated through computer applications. These applications can be used to predict and analyze the future. To achieve this phenomenon, we train the machines to read data and accordingly predict the future which is called as Machine Learning. Machine learning is done by training the machine to react to different data inputs. The paper focuses on a brief overview of machine learning and few machine learning algorithms and techniques as well as risks and applications of it.*

Key Words: Artificial Intelligence, Machine Learning, Algorithms, Neural Networks.

1. INTRODUCTION

Machine Learning (ML) is a subset of Artificial Intelligence (AI). Using Machine Learning we can make applications learn from experience in the same way as human do. When data is fed into these applications, they learn grow and change according to experience. This is done by using algorithms that learn from data in an iterative process. Applications use pattern recognition to respond to various data. Machine learning is the ability of an applications to react to new data using iterations. Machine learning algorithms learn how to predict outputs based on previous examples of relationships between input data and outputs which is called as training data. A model of the relationship between inputs and outputs is gradually improved by testing its predictions and correcting when wrong. Machine learning is a set of computerised techniques for recognising patterns in data. It is a way of generating something like the line of best fit. It's useful to automate this process when the data has many features and is very complex.

2. TYPES OF MACHINE LEARNING

2.1 Supervised Learning

Supervised learning requires a training data setwithlabelled data, or data with a known output value (e.g. rural/not rural or house price). Classification and regression problems are solved through supervised learning.

2.2 Unsupervised Learning

Unsupervised learning techniquesdon't use a trainingset and find patterns or structure in the data by themselves. Clustering problems can be solved with an unsupervised approach.

2.3 Semi-Supervised Learning

Semi-supervised learning uses mainly unlabeled and a small amount of labelled input data. Using a small amount of labelled data can greatly increase the efficiency of unsupervised learning tasks. The model must learn the structure to organize the data as well as make predictions.

2.4 Reinforcement Learning

Reinforcement learning uses input data from the environment as a stimulus for how the model should react. Feedback is not generated through a training process like supervised learning but as rewards or penalties in the environment. This type of process is used in robot control.

3. HOW DOES MACHINE LEARNING WORK?

Machine Learning process starts with feeding data to an algorithm. This is also called as training thealgorithm. To test whether an algorithm is working, data is fed into the algorithm and results are checked. If desired results are not achieved, then the algorithm is re-trained and tested again. This process is repeated until we achieve the desired results. This helps the machine learning algorithm to learn and give the desired output as well as increases the accuracy of the result.

4. HOW DOES MACHINE LEARNING WORK?

In the process of machine learning, the quality of information that external environment provides to the system is the primary factor. The external environment is outside information set that delivers itself in some form, it represents sources of outside information; Learning is the process that processes the outside information to knowledge first it obtains the information of outside environment and then processes the information to knowledge, and puts this knowledge into the repository; Repository stores many general principles that guide a part of the implementation action, due to environment provides all kinds of information for learning system, the quality of information impacts

directly on learning realization whether easy or disorderly. Repository is the second factor that impacts the design of learning system. The expression of knowledge is varied, such as, eigenvector, logic statements of the first order, production rules, semantic networks and frameworks and so on, these fashions of expression each has its strong point. Consider four aspects when to choose: strong in expression, easy to infer, easy to modify repository, the knowledge is easy to expand. The implementation is the process that uses the knowledge of repository to complete a certain task, and to feed back the information which obtained in the progress of completing the task to the learning, and to guide further study.

5. MACHINE LEARNING ALGORITHMS

Algorithms are often grouped by similarity in terms of their function (how they work). For example, tree-based methods, and neural network inspired methods. I think this is the most useful way to group algorithms and it is the approach we will use here. This is a useful grouping method, but it is not perfect. There are still algorithms that could just as easily fit into multiple categories like Learning Vector Quantization that is both a neural network inspired method and an instance-based method. There are also categories that have the same name that describe the problem and the class of algorithm such as Regression and Clustering.

5.1 Decision Tree Based Classification

Decision tree algorithm mainly used to construct a training/classification/regression model in the form of a tree structure (root, branch and leaf), which is based on (inferred from) previous data to classify/predict class or target variables of future/new data with the help of decision rules or decision trees. Decision Trees can be used for numericalas well as categorical data. The algorithm works on greedy-search technology starting from top to bottom.

Advantages:

- Easy to implement
- Can classify and predict categorical as well as numerical data.
- Less data pre-processing
- Statistical test can be done to validate the tree model. Resembles human decision-making technology.
- Tree structure is easily understandable through visualization

Limitations:

- Low Prediction Accuracy
- Complexity in calculations if class labels are huge
- Need of redrawing for every addition of data to the data set.
- Probability of over-fitting in the decision tree is high.

Applications:

- Agriculture
- Medicine
- Financial analysis

5.2 Support Vector Machines

Supervised learning algorithm. It used to perform regression, classification, and outlier detection of data. The main objective is to find a hyper plane that best divides the two classes. Based on this hyper plane, the new data is best classified to which class it belongs to. Two rules must be followed while drawing a hyper plane. First is that the hyper plane must separate the two classes and maximum-margin hyper plane should be chosen as best hyper separator. There are two types of SVM Classifiers: a) Linear SVM Classifier b) Non-Linear SVM Classifier

Advantages:

- Robust Classifier for prediction problem
- Efficient if the number of dimensions is greater than the number of samples
- Suitable for high dimensional data spaces
- Memory efficient.

Disadvantages:

- Very slow in test phase
- Not suitable for large and noisy data sets
- Classification error percentage will be increased if wrong kernel is selected
- Memory consumption is high

Applications:

- Image classification
- Bioinformatics
- Face detection

5.3 K-Nearest Neighbors

K-N N is a supervised classifier. It is a best choice for the classification kind of problems. To predict the target label of a new test data, KNN finds the distance of nearest training data class labels with a new test data point in the presence of K value. Then counts the number of very closest data points using K value and concludes the new test data class label. To calculate the number of nearest training data points distance, KNN uses K variable value between 0 to 1o normally. Among the popular distance functions like Euclidean distance, Manhattan distance, Minkowski distance and Hamming distance, Euclidean distance function is used for continuous variables and Hamming distance function is used for categorical variables.

Advantages:

- Flexible with attributes and distance functions
- Supports multi class

dataDisadvantages:

- Finding suitable K value is difficult
- Needs large sample for high accuracy
- Requires large storage space.

5.4 Naïve Bayes Algorithm

The algorithm performs classification tasks in the field of machine learning. It can do classification very well on the dataset even it has huge records with multi class and binary class classification problems. The main application of Naive Bayes is text analysis and Natural Language Processing. Understanding of Bayes theorem will help to understand (work with) Naive Bayes algorithm efficiently. Bayes theorem is used to combine the multiple classification algorithms to form Naive Bayes classifier with a common principle. Bayes theorem works based on conditional probability. Conditional probability means, an event will occur with conditioned (based) on an event already occurred.

Types of Naïve Bayes:

a) Gaussian Naïve Bayes

b) Multinomial Naïve Bayes

c) Bernoulli Naïve Bayes

Advantages:

- Fast
- Scalable
- Efficient for binary and multinomial distributed attribute values

Disadvantages:

- Not suitable for regression problems
- Cannot find relationships among attributes

Applications:

- Text, e-mail, and symbol analysis
- Recommendation Systems

5.5 Linear Regression

It finds the relationship between an independent (predictor (X)) and a dependent (criterion (Y)) variable to predict the future values of the dependent variable. Simple regression uses one independent variable and multiple regressions use two or more independent variables to predict the future. Dependent variable has a continuous and independent variable has discrete or continuous values. There are two kinds of regression models. One is linear and other one is non-linear. The linear regression model uses straight line and non-linear regression model uses curved line relationships between dependent and independent variables.

Advantages:

- Shows relationship between dependent and independent variables
- Simple and easy to understand

Disadvantages:

- Not applicable for non-linear data
- Only predicts numerical output
- Data must be independent

Applications:

- Observational Astronomy
- Finance

6. DETAILED COMPARISON OF THE ALGORITHMS

Learning Method	Parameter Estimation Algorithm	Model Complexity Reduction
Gaussian Naïve Bayes	Estimate $\hat{\mu}$, $\hat{\sigma}2$, and P(Y) using maximum likelihood	Place prior on parameters and use MAP estimator
Logistic Regression	No closed form estimates. Optimize objective function using gradient descent	L2 regularization
Decision Trees	Many algorithms: ID3, CART, C4.5	Prune tree or limit tree depth
K-Nearest Neighbors	Must store all training data to classify new points. Choose K using cross validation.	Increase K
Support Vector Machines (with slack variables, no kernel)	Solve quadratic program to find boundary that maximizes margin	Reduce C

Learning Method	Generative or Discriminative	Loss Function	Decision Boundary
Gaussian Naïve Bayes	Generative	– log P(X, Y)	Equal variance: linear boundary. Unequal variance: quadratic boundary
Logistic Regression	Discriminative	– log P(Y |X)	Linear
Decision Trees	Discriminative	Either – log P(Y |X) or zero-one loss	Axis-aligned partition of feature space
K-Nearest Neighbors	Discriminative	zero-one loss	Arbitrarily complicated
Support Vector Machines (with slack variables, no kernel)	Discriminative	hinge loss: |1−y(wT x)|+	linear (depends on kernel)

7. RISKS IN MACHINE LEARNING

7.1 Data Poisoning

Data play an outsized role in the security of an ML system. That's because an ML system learns to do what it does directly from data. If an attacker can intentionally manipulate the data being used by an ML system in a coordinated fashion, the entire system can be compromised. Data poisoning attacks require special attention. ML engineers should consider what fraction of the training data an attacker can control and to what extent.

7.2 Data confidentiality

Data protection is difficult enough without throwing ML into the mix. One unique challenge in ML is protecting sensitive or confidential data that, through training, are built right into a model. Subtle but effective extraction attacks against an ML system's data are an important category of risk.

7.3 Online system manipulation

An ML system is said to be online when it continues to learn during operational use, modifying its behavior over time. In this case, a clever attacker can nudge the still- learning system in the wrong direction on purpose through system input and slowly retrain the ML system to do the incorrect thing. Note that such an attack can be both subtle and reasonably easy to carry out. This risk is complex, demanding that ML engineers consider data provenance, algorithm choice, and system operations to properly address it.

7.4 Over-fitting

Overfitting happens when a model learns the detail and noise in the training data to the extent that it negatively impacts the performance of the model on new data. This means that the noise or random fluctuations in the training data is picked up and learned as concepts by the model. The problem is that these concepts do not apply to new data and negatively impact the model's ability to generalize. Overfitting is more likely with nonparametric and nonlinear models that have more flexibility when learning a target function. As such, many nonparametric machine learning algorithms also include parameters or techniques to limit and constrain how much detail the model learns.

8. ADVANTAGES OF MACHINE LEARNING

8.1 Easily identifies trends and patterns

Machine Learning can review large volumes of data and discover specific trends and patterns that would not be apparent to humans. For instance, for an ecommerce website like Amazon, it serves to understand the browsing behaviors and purchase histories of its users to help cater to the right products, deals, and reminders relevant to them. It uses the results to reveal relevant advertisements to them.

8.2 Continuous improvements

As Machine Learning algorithms gain experience, they keep improving in accuracy and efficiency. This lets them make better decisions. Say you need to make a weather forecast model. As the amount of data you have keeps growing, your algorithms learn to make more accurate predictions faster

8.3 No human intervention needed (automation) With ML, you don't need to make the changes manually, since it means giving machines the ability to learn, it lets them make predictions and also improve the algorithms on their own. A common example of this is anti-virus software's; they learn to filter new threats as they are recognized. ML is also good at recognizing spam.

9. APPLICATIONS OF MACHINE LEARNING

9.1 Driverless Cars

Machine learning algorithms are an integral part of driverless cars and will have an increasingly important role in their operational ability. These learning systems are broadly used for tasks such as image recognition or scheduling but learning in noisy real–world environments is difficult. Trying to record, evaluate and combine lots of different types of data is very challenging. The solution so far has been to try and capture as many of these rare events as possible through extensive testing on the road. Using this data, algorithms are used to devise responses which are then tested in simulations.

9.2 Medicine, Healthcare

When it comes to diagnosis or decision making, machine learning algorithms are not a good replacement for clinicians - at least in most situations. A good diagnosis must take into account structured (e.g. diagnosis codes, medications), unstructured data (clinical notes), image data (X-rays) and even subtle visual cues from the patient (do they look ill, how did they answer family history questions) in a very short time frame.

9.3 Machine learning in Government

Machine learning is currently exploited by a handful of people in government such as the Government Digital Services (GDS) who are using it to predict page views to do anomaly detection11 or the HMRC who are using clustering techniques to segment VAT customers. Uptake is still limited and there is a great deal of untapped potential. GDS have so far focused on demonstrating the

capabilities of machine learning algorithms on a number of products and prototype services. One of the first steps to increasing exploitation of these learning systems is to develop a 'data first' mind-set at a much earlier stage in the policy process.

10. CONCLUSION

This paper introduces the concept of machine learning, the basic model, and its application in various field along with its advantages and disadvantages. It also reviews various machine learning techniques and tools such as classification and prediction techniques, including objective, working procedure, advantages, limitations, Real-time Applications, and Implementation tools. The emerging trends in Artificial Intelligence and machine learning need the strong fundamentals of above-mentioned methods and will be useful in interdisciplinary areas also.

REFERENCES

[1] Jafar Tanha, "Semi-supervised self-training for decision tree classifiers", International Journal of Machine Learning and Cybernetics, Volume 8, Issue 1, Page No's: 355-370, January, 2015.
[2] Khadim D, Fleur M and Gayo D, "Large scale biomedical texts classification: a k-NN and an ESA-based approaches",
Journal of Biomedical Semantics, 7:40, June, 2016
[3] Hong R, H. M. Wang and Jian L, "PrivacyPreserving kNearest Neighbor Computation in Multiple Cloud Environments, IEEE Access, ISSN: 2169-3536, Volume 4, Page No's: 9589-9603, December, 2016.
[4] L. Jiang, C. Li, "Deep feature weighting for naive Bayes and its application to text classification", Journal of Engineering Applications of Artificial Intelligence, Volume 52, Page No's: 26-39, June, 2016.
[5] Ahmed M, Alison H, "Modeling built-up expansion and densification with multinomial logistic regression, cellular automata and genetic algorithm", Volume 67, Page No's: 147- 156, January, 2018
[6] T. Razzaghi, Oleg R, "Multilevel Weighted Support Vector Machine for Classification on Healthcare Data with Missing Values", PLUS ONE, Page No's:1-18, May 2016
[7] Hui L, D. Pi, "Integrative Method Based on Linear Regression for the Prediction of Zinc binding Sites in Proteins", IEEE Access, Volume 5, Page No's: 14647-14656, August, 2017.
[8] L. Wang, D. Wang, "Intelligent CFAR Detector Based on Support Vector Machine", IEEE Access, Volume 5, Page No's: 26965-26972, December, 2017.
[9] Enrico R, Michel L, "The Counter, a Frequency Counter Based on the Linear Regression", IEEE Transactions on Ultrasonics, Ferroelectrics, Volume 63, Issue 7, Page No's: 961-969, July, 2016.
[10] J. Han, M. Kamber and J. Pei, "Data Mining: Concepts and Techniques", 3rd Edition, MK Series, 2012.

[11] http://www.maxlittle.net/publications/index.php 10.
[12] https://gdsdata.blog.gov.uk/2014/08/15/anomalydetec tion-a-machine-learning-approach/ http://www.ons.gov.uk/ons/.../mwp1-ons-innovationlaboratories.pdf
[13] https://berryvilleiml.com/ results/ara.pdf
[14] X. Yu, P. He, Q. Zhu, and X. Li, "Adversarial examples: Attacks and defenses for deep learning," IEEE Trans. Neural Netw. Learn. Syst., vol. 30, no. 9, pp. 2805–2824, 2019. doi: 10.1109/TNNLS.2018.2886017
[15] https://data-flair.training/blogs/advantages-anddisadvantages-of-machine-learning/
[16]https://www.burtchworks.com/2018/06/12/2018machine-learning-flash-survey-results/