Machine Learning & its Classification Techniques

Atul B.Kathole, Prasad S. Halgaonkar, Ashvini A.Nikhade

Abstract: *Recent developments in information systems as well as computerization of business processes by organizations have led to a faster, easier and more accurate data analysis and accuracy. Data mining and machine learning techniques have been used increasingly in the analysis of data in various fields ranging from medicine to organization, education and energy applications. Machine learning techniques make it possible to deduct meaningful further information from those data processed by data mining. Such meaningful and significant information helps organizations to establish their future policies. This study applies classification machine learning techniques also survey of active learning regarding selection methods, query strategies, applications.*

Keyword: *Machine Learning, Data Mining, Classification Techniques.*

I. INTRODUCTION:

Machine Learning is an approach or subset of Artificial Intelligence that is based on the idea that machines can be given access to data along with the ability to learn from it. It learns from examples and experience, without being explicitly programmed. Instead of writing code, you feed data to the generic algorithm, and it builds logic based on the data given. In a very layman manner, Machine Learning (ML) can be explained as automating and improving the learning process of computers based on their experiences without being actually programmed i.e. without any human assistance. The process starts with feeding good quality data and then training our machines (computers) by building machine learning models using the data and different algorithms. The choice of algorithms depends on what type of data do we have and what kind of task we are trying to automate.

Revised Manuscript Received on July 25, 2019.

Atul B.Kathole, Faculty,Department of Computer Engineering, Zeal College of Engineering and Research, Narhe, Pune, India.

Dr.Prasad S. Halgaonkar, Faculty, Department of Computer Engineering, Zeal College of Engineering and Research, Narhe, Pune, India.

Ashvini A.Nikhade, Faculty, Department of Computer Engineering, Zeal College of Engineering and Research, Narhe, Pune, India.

atul.kathole@zealeducation.com,prasad.halgaonkar@zealeducation.com, ashvini.nikhade@zealeducation.com

Data Science and Machine Learning

- Data Science and Machine Learning go hand in hand. Data Science helps evaluate data for Machine Learning algorithms
- Data science is the use of statistical methods to find patterns in the data.
- Statistical machine learning uses the same math and techniques as data science.
- These techniques are integrated into algorithms that learn and improve on their own.
- Machine Learning facilitates Artificial Intelligence as it enables machines to learn from the patterns in data.

Difference between Traditional Programming & Machine Learning

- **Traditional Programming:** We feed in DATA (Input) + PROGRAM (logic), run it on machine and get output.
- **Machine Learning :** We feed in DATA(Input) + Output, run it on machine during training and the machine creates its own program(logic), which can be evaluated while testing.

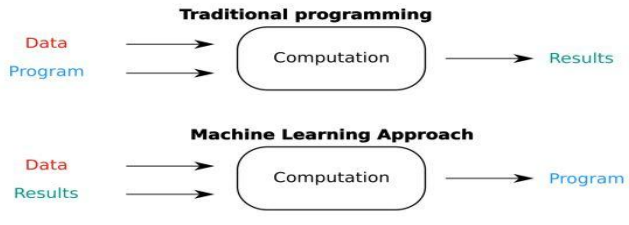

Figure 01: Difference between Traditional Programming & Machine Learning

What does exactly learning means for a computer?

A computer is said to be learning from **Experiences** with respect to some class of **Tasks**, if its performance in a given Task improves with the Experience. A computer program is said to learn from experience **E** with respect to some class of tasks **T** and performance measure **P**, if its performance at tasks in **T**, as measured by **P**,

Retrieval Number: I30280789S319/2019©BEIESP
DOI: 10.35940/ijitee.I3028.0789S319
Published By:
Blue Eyes Intelligence Engineering
& Sciences Publication

improves with experience **E**

Need of Machine Learning

Machine Learning is a field which is raised out of Artificial Intelligence (AI). Applying AI, we wanted to build better and intelligent machines. But except for few mere tasks such as finding the shortest path between point A and B, we were unable to program more complex and constantly evolving challenges.

There was a realization that the only way to be able to achieve this task was to let machine learn from itself. This sounds similar to a child learning from its self. So machine learning was developed as a new capability for computers. And now machine learning is present in so many segments of technology, that we don't even realize it while using it. Finding patterns in data on planet earth is possible only for human brains. The data being very massive, the time taken to compute is increased, and this is where Machine Learning comes into action, to help people with large data in minimum time. If big data and cloud computing are gaining importance for their contributions, machine learning as technology helps analyze those big chunks of data, easing the task of data scientists in an automated process and gaining equal importance and recognition. The techniques we use for data mining have been around for many years, but they were not effective as they did not have the competitive power to run the algorithms. If you run deep learning with access to better data, the output we get will lead to dramatic breakthroughs which is machine learning.

II. TECHNIQUES OF MACHINE LEARNING

The Machine Learning has three main Techniques those are as follow,

- Supervised learning
- Unsupervised learning
- Semi-supervised and Reinforcement learning

i) Supervised Learning:

Supervised Learning is a type of Machine Learning used to learn models from labeled training data. It allows us to predict the output for future or unseen data.

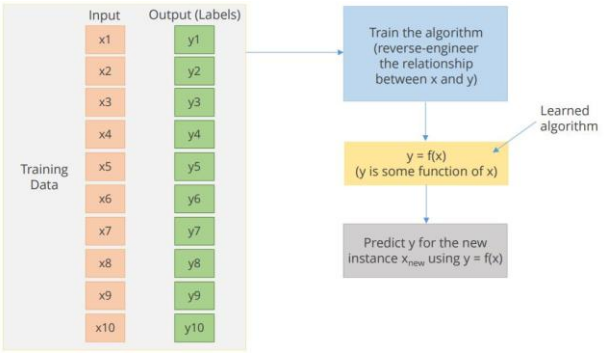

Figure 02 : Algorithm of Supervised Learning

Supervised Learning Flow

- Data Preparation
- Training Step
- Evaluation or Test Step
- Production Deployment

Testing the Algorithm

Given below are the steps for testing the algorithm of Supervised Learning.

1. Once the algorithm is trained, test it with test data (a set of data instances that do not appear in the training set).
2. A well-trained algorithm can predict well for new test data.
3. If the learning is poor, we have an underfit situation. The algorithm will not work well on test data. Retraining may be needed to find a better fit.
4. If learning on training data is too intensive, it may lead to overfitting – a situation where the algorithm is not able to handle new testing data that it has not seen before. The technique to keep data generic is called regularization.

Types of Supervised Learning

Given below are 2 types of Supervised Learning.

- Classification
- Regression

Classification Supervised Learning

Let us look at the classifications of Supervised learning.

- Answers "What class?"
- Applied when the output has finite and discrete values

Retrieval Number: I30280789S319/2019©BEIESP
DOI: 10.35940/ijitee.I3028.0789S319
Published By:
Blue Eyes Intelligence Engineering
& Sciences Publication

Regression Supervised Learning

Given below are some elements of Regression Supervised learning.

- Answers "How much?"
- Applied when the output is a continuous number
- A simple regression algorithm: y = wx + b. Example: the relationship between environmental temperature (y) and humidity levels (x)

ii) Unsupervised Learning:

Unsupervised learning is a term used for Hebbian learning, associated to learning without a teacher, also known as self-organizationand a method of modelling the probability density of inputs.

The unsupervised learning to create clusters of heavenly bodies, with each cluster containing objects of a similar nature. Unsupervised Learning is a subset of Machine Learning used to extract inferences from datasets that consist of input data without labeled responses.

Compared to supervised learning where training data is labeled with the appropriate classifications, models using unsupervised learning must learn relationships between elements in a data set and classify the raw data without "help." This hunt for relationships can take many different algorithmic forms, but all models have the same goal of mimicking human logic by searching for indirect hidden structures, patterns or features to analyze new data.

Types of Unsupervised Learning

There are three types of Unsupervised Learning are:

- Clustering
- Visualization Algorithms
- Anomaly Detection

Clustering

The most common unsupervised learning method is cluster analysis. It is used to find data clusters so that each cluster has the most closely matched data.

Visualization Algorithms

Visualization algorithms are unsupervised learning algorithms that accept unlabeled data and display this data in an intuitive 2D or 3D format. The data is separated into somewhat clear clusters to aid understanding.

Anomaly Detection

This algorithm detects anomalies in data without any prior training. It can detect suspicious credit card transactions and differentiate a criminal from a set of people.

iii) Semi-Supervised Learning

It is a hybrid approach (combination of Supervised and Unsupervised Learning) with some labeled and some non-labeled data.

Semi-supervised learning is a class of machine learning tasks and techniques that also make use of unlabeled data for training – typically a small amount of labeled data with a large amount of unlabeled data. Semi-supervised learning falls between unsupervised learning (without any labeled training data) and supervised learning (with completely labeled training data). Many machine-learning researchers have found that unlabeled data, when used in conjunction with a small amount of labeled data, can produce considerable improvement in learning accuracy over unsupervised learning (where no data is labeled), but without the time and costs needed for supervised learning (where all data is labeled).The acquisition of labeled data for a learning problem often requires a skilled human agent (e.g. to transcribe an audio segment) or a physical experiment (e.g. determining the 3D structure of a protein or determining whether there is oil at a particular location). The cost associated with the labeling process thus may render a fully labeled training set infeasible, whereas acquisition of unlabeled data is relatively inexpensive. In such situations, semi-supervised learning can be of great practical value. Semi-supervised learning is also of theoretical interest in machine learning and as a model for human learning.

III. DATA PREPROCESSING IN MACHINE LEARNING

Here, we will Discuss Data Preprocessing, Feature Scaling, and Feature Engineering in detail

Objectives

- Recognize the importance of data preparation in Machine Learning
- Identify the meaning and aspects of feature engineering
- Standardize data features with feature scaling
- Analyze datasets and its examples
- Explain dimensionality reduction with Principal Component Analysis (PCA)

Data Preparation in Machine Learning

Retrieval Number: I30280789S319/2019©BEIESP
DOI: 10.35940/ijitee.I3028.0789S319
Published By:
Blue Eyes Intelligence Engineering
& Sciences Publication

A quick brief of Data Preparation in Machine Learning is mentioned below.

- Machine Learning depends largely on test data.
- Data preparation is a crucial step to make it suitable for ML.
- A large amount of data is generally required for the most common forms of ML.
- Data preparation involves data selection, filtering, transformation, etc.

Data Preparation Process

The process of preparing data for Machine Learning algorithm comprises the following:

- Data Selection
- Data Preprocessing
- Data Transformation

Data Selection

Steps involved in Data Selection involves:

- There is a vast volume, variety, and velocity of available data for a Machine Learning problem.
- This step involves selecting only a subset of the available data.
- The selected sample must be an accurate representation of the entire population.
- Some data can be derived or simulated from the available data if required.
- Data not relevant to the problem at hand can be excluded.

Data Preprocessing

Let's understand Data Preprocessing in detail below.
After the data has been selected, it needs to be preprocessed using the given steps:

1. Formatting the data to make it suitable for ML (structured format)
2. Cleaning the data to remove incomplete variables
3. Sampling the data further to reduce running times for algorithms and memory requirements.

Data cleaning at this stage involves filtering it based on the following variables:
Insufficient Data: The amount of data required for ML algorithms can vary from thousands to millions, depending upon the complexity of the problem and the chosen algorithm.

Non-Representative Data
The sample selected must be an exact representation of the entire data, as non-representative data might train an algorithm such that it won't generalize well on new test data.

Substandard Data
Outliers, errors, and noise can be eliminated to get a better fitment of the model. Missing features such as age for 10% of the audience may be ignored completely, or an average value can be assumed for the missing component.
Selecting the right size of the sample is a key step in data preparation. Samples that are too large or too small might give skewed results.

Sampling Noise
Smaller samples cause sampling noise since they get trained on non-representative data. For example, checking voter sentiment from a very small subset of voters.

Sampling Bias
Larger samples work well as long as there is no sampling bias, that is, then the right data is picked. For example, sampling bias would occur when checking voter sentiment only for the technically sound subset of voters, while ignoring others.

Data Transformation

The selected and preprocessed data is transformed using one or more of the following methods:

1. Scaling: It involves selecting the right feature scaling for the selected and preprocessed data.
2. Aggregation: This is the last step to collate a bunch of data features into a single one.

IV. TYPES OF DATA

Labeled Data or Training Data

- It is also known as marked (with values) data.
- It assists in learning and forming a predictive hypothesis for future data. It is used to arrive at a formula to predict future behavior.
- Typically 80% of available labeled data is marked for training.

Unlabeled Data

- Data which is not marked and needs real-time unsupervised learning is categorized as unlabelled data.

Retrieval Number: I30280789S319/2019©BEIESP
DOI: 10.35940/ijitee.I3028.0789S319
Published By:
Blue Eyes Intelligence Engineering
& Sciences Publication

Test Data

- Data provided to test a hypothesis created via prior learning is known as test data.
- Typically 20% of labeled data is reserved for the test.

Validation data

It is a dataset used to retest the hypothesis (in case the algorithm got over fitted to even the test data due to multiple attempts at testing).

The illustration given below depicts how total available labeled data may be segregated into the training dataset, test dataset, and validation dataset.

V. CONCLUSION:

Today, it is inevitable to consider and use data mining in view of the ever-increasing amount of computerized business processes and the huge amount of data to be analyzed in parallel. It is possible to make accurate estimations or predictions about future results through applying machine learning techniques to the data made available for analysis via data mining. This study used in various classification techniques to a group of individuals who are in the process. Since using machine learning techniques in classification studies results in accurate outcomes accompanied with significant saving in terms of time and cost, it is highly recommended to make use of machine learning techniques in Data Processing. It is considered that this study will be useful for organizations and individuals operating in all areas of work that have gone for computerization of their business processes.

REFERENCES:

1. J. Qi, P. Yang, G. Min, O. Amft, F. Dong, and L. Xu, "Advanced Internet of Things for personalised healthcare systems: A survey," Pervasive and Mobile Computing, vol. 41, pp. 132-149, 2017.
2. T. Pawar, N. Anantakrishnan, S. Chaudhuri, and S. P. Duttagupta, "Impact of ambulation in wearable-ECG," Annals of Biomedical Engineering, vol. 36, no. 9, pp. 1547-1557, 2008.
3. L.-J. Kau and C.-S. Chen, "A smart phone-based pocket fall accident detection, positioning, and rescue system," IEEE journal of biomedical and health informatics, vol. 19, no. 1, pp. 44-56, 2015.
4. M. Ermes, J. Pärkkä, J. Mäntyjärvi, and I. Korhonen, "Detection of daily activities and sports with wearable sensors in controlled and uncontrolled conditions," IEEE transactions on information technology in biomedicine, vol. 12, no. 1, pp. 20-26, 2008.
5. D. Riaño et al., "An ontology-based personalization of healthcare knowledge to support clinical decisions for chronically ill patients," Journal of biomedical informatics, vol. 45, no. 3, pp. 429-446, 2012.
6. M. S. Organization. (2018, Nov 2). Top 83 AI startups in Healthcare. Available: http://www.medicalstartups.org/top/ai.
7. Y. Gordienko et al., "Deep learning with lung segmentation and bone shadow exclusion techniques for chest x-ray analysis of lung cancer," in International Conference on Theory and Applications of Fuzzy Systems and Soft Computing, 2018, pp. 638-647: Springer.
8. M. T. Brown and J. K. Bussell, "Medication adherence: WHO cares?," Mayo Clinic proceedings, vol. 86, no. 4, pp. 304-314, 2011.
9. K. Teng. (2012, May). What Is Personalized Healthcare? From patients to medications, one size does not fit all. Available: https://health.clevelandclinic.org/what-is-personalizedhealthcare.
10. A. Ara and A. Ara, "Case study: Integrating IoT, streaming analytics and machine learning to improve intelligent diabetes management system," in 2017 International Conference on Energy, Communication, Data Analytics and Soft Computing (ICECDS), 2017, pp. 3179-3182.
11. R. Vargheese and Y. Viniotis, "Influencing data availability in IoT enabled cloud based e-health in a 30 day readmission context," in 10th IEEE International Conference on Collaborative Computing: Networking, Applications and Worksharing, 2014, pp. 475-480.
12. M. D. Naylor, L. H. Aiken, E. T. Kurtzman, D. M. Olds, and K. B. Hirschman, "The importance of transitional care in achieving health reform," Health affairs, vol. 30, no. 4, pp. 746-754, 2011.
13. S. Tilson and G. J. Hoffman, "Addressing Medicare hospital readmissions," Congressional Research Service, 2012.
14. S. Shahrestani, "Assistive IoT: Deployment Scenarios and Challenges," in Internet of Things and Smart Environments: Springer, 2017, pp. 75-95.
15. M. A. Hanson et al., "Body Area Sensor Networks: Challenges and Opportunities," Computer, vol. 42, no. 1, pp. 58-65, 2009.
16. M. J. Rothman, S. I. Rothman, and J. Beals IV, "Development and validation of a continuous measure of patient condition using the Electronic Medical Record," Journal of biomedical informatics, vol. 46, no. 5, pp. 837-848, 2013.

Retrieval Number: I30280789S319/2019©BEIESP
DOI: 10.35940/ijitee.I3028.0789S319

Published By:
Blue Eyes Intelligence Engineering
& Sciences Publication