MACHINE LEARNING: SURVEY, TYPES AND CHALLENGES

Pallavi D. Bhalekar[1], Dr. M. Z. Shaikh[2]

[1]Bharati Vidyapeeth College of Engineering, University of Mumbai
[2] Bharati Vidyapeeth College of Engineering, University of Mumbai

---***---

Abstract - *Now a day's AI is popular topic in an industry. Machine learning is core part of AI. In this paper we try to do study of machine learning and its types. Which algorithms are used into machine learning as per its categories Feature selection method is studied and used to provide the accurate model. Reducing unnecessary features from given dataset the feature selection method is used. Predictive model is used to predict what will be the output in future if we provide particular input. Predictive model is part of machine learning and it's learning from the previous behavior of system. In this paper explained the different types of machine learning algorithms like that SVM, KNN, naive byes, decision tree algorithm etc. These algorithms are divided into two groups as supervised machine learning and unsupervised machine learning algorithm.*

Key Words: Machine learning, types of machine learning, prediction model, feature selection.

1.INTRODUCTION

Machine learning is one of the foundations of countless applications like that web search, product recommendation, speech recognition, robotics, social networks, e-commerce and many more. Machine learning is the part of artificial intelligence that provides the system ability to automatically improve from our experience without being explicitly programmed. The system learning process will start from the prior knowledge of the system like that by observing pattern, instructions, direct knowledge etc. The main aim of machine learning is computers had to work automatically without any human interaction and have to adjust actions accordingly as per the situation.
Machine learning aids in solving business problem by adding large amount of data. In organizations use machine learning have to scalable data prepetition capabilities. The main advantages of machine learning is that quickly produce models with large amount of data and result will deliver very fast as well as accurate result.

From last 21st century, every business is realizing that machine learning will increase calculation potential. After that many projects are launched like that Google Brain, Deep face, Deep mind, Amazon machine learning platform U-net and many more. Google Brain is deep learning AI research team at google. Many projects are developed under google brain. Like that devised encryption system, Image enhancement, Robotics, Google translates etc. Deep face is deep neural network created by Facebook. The aim of this application is recognize people with same claim as human can recognize. Deep mind is totally related with video game. Amazon Machine Learning is core part of the Amazon Machine web service. It create platform for big companies which is getting to involve into machine learning. It drives many internal systems like that search recommendation, Alexa, Amazon Go etc. U-net is used in CNN (Canvolutional Neural Network) architecture specialized in biomedical image segmentation. After passing sampling data through CNN architecture obtained a labeled output that can be classified.
As per types of machine learning model the algorithms are used as linear regression. Linear regression is finding the line that works best between given set of points. We will understand it by giving example. Like there are two houses. First house is small with prize 30 lac. And second house is of 1 cr. In prize and we have to guess the size of medium size house. For this we will plot the graph as size of house on x- axis and prize of house on y-axis. We have to find best solution for medium house the prize in between 30 lac to 1 cr. But what is the exact prize. It may be 50 lac. 70 lac. Or 90 lac. We will draw the line from given set of point and get exact prize of the house as 70 lac. This method is called as linear regression. Next algorithm is gradient descent algorithm. The aim of this algorithm is Square of error minimization to get best line fit. Suppose we have three points which is in scattered format. How to find best solution for these three points as by calculating error rate of drawn line. Adjust line fit up to error will have to minimize. Third algorithm is naive byes algorithm. We will take example of e-mail spam detection. This algorithm is worked on previous information or data. Suppose the history of the given e mail spam is the message contain with cheap word those by looking at history we will guess the the spam message. As calculating probability of how many times that word is occurred in spam message or in non-spam message. Then as per algorithm decides that mail is spam or not. Next algorithm is decision tree algorithm. For this we will take example of

recommendation engine as we have recommends apps. For this we have table with three columns like that gender age and app. Age which is less than 20 years old both male and females download the games. We can create tree as two child nodes like that less than 20 years and greater than 20 years. Like wise decision tree can found.

Types of machine learning

1.1 Supervised machine learning: Supervised learning is enables the model to predict future outcomes after they are trained based on past information. This means machine will learn from labeled data. Example of supervised learning is suppose we have to identify car; and we have data related with vehicle like that tires, air suspension, steering, engine, headlights likewise but we have to identify only cars then some extra feature like setting capacity, height of vehicle, model of the vehicle from these information we will identify the cars. This information is labeled if it has these features then probability that it must be car. In supervised learning we have to map our output. These past data or information are known as dataset. Dataset is divided into two groups like training set and testing set. Training set contains with all features information. From above explanation we understood that in supervised learning we are trying to infer function from the training datasets. Training model is worked as per learning algorithm.

Supervised learning is used to solve two types of problems 1] classification 2] Regression

Classification and regression both are related with prediction. But difference between those two concept is regression predicts values from continues set and classification predicts belonging to the class.

1.1.1] classification: The concept of categorizing data is based on training with set of the data so that machine can essentially learn boundaries that separate categories of data. Classification is used to predict discrete values as which class is data point of given dataset. In classification problem data must be divided into one of two or more classes. Classification algorithm may predict continues values but main requirement of classification is that continues value is in the form of probability for a class label. Classification can be evaluated using accuracy. Algorithms used into classifications are logistic regression, Decision tree, k nearest neighbors etc

1.1.2] Regression: Regression means to predict the output values using training data. Regression is used to predict continuous quantity. A regression algorithm may predict discrete values, but the discrete value is in the form of integer quantity. Regression prediction can be evaluated using root mean squared error. Algorithm used into regression are Random forest, Linear regression etc.

1.2 Unsupervised machine learning: Unsupervised learning is the training to the machine without any labeled data or information and allows to the algorithm to act on that information without any prior knowledge. In unsupervised learning we have to find out hidden pattern from given data unsupervised learning is very helpful in exploratory data analysis (EDA) because it can automatically find out hidden pattern from the given data. As per given pattern, training data contains with input values without any corresponding output or target values. The goal of unsupervised learning is cluster the given input data into characteristically different groups. Given data can be clustered into different different groups depending upon way of clustering. Clustering is related to find similar groups or similar entities within large sensional data. We will understand the concept of unsupervised learning with example. Suppose we are watching news online by searching on google like news.google.com there are number of hyperlinks provided to us. If we select one news hyperlink it may contains with other link that news are divided into groups like politics, weather, fashion likewise in one hyperlink contains with number of links. If we clicked on all links they will provide different web pages of that particular news. That single hyperlink consists of number of sub-hyperlinks this method is nothing but clustering. Clustering is groups of similar data.

1.2.1] Clustering: In supervised learning data is labeled data. That's why processing on labeled data is very easy task. But unsupervised model is worked on the unlabelled data. We have to convert that unlabelled data into labeled data the clustering technique is used. Therefore the important part of unsupervised learning is clustering. Clustering is nothing but process of collecting unlabelled data into similar groups. And remaining data into different groups. There are mainly three techniques of clustering those are 1] Hierarchical 2] Partition 3] Bayesian

Hierarchical clustering technique: In this technique clustered are defined in the form of tree type structure. It may be used bottom up approach or top down approach. Bottom up approach is also known as agglomerative technique. And top down technique is known as divisive technique. In bottom up approach each observation is allot to its own cluster and then similarity is computed. In top down all the observed data is allotted to single cluster and then it divided into more than two similar clusters.

Partition clustering technique: In this clustering technique each cluster is divided into k clusters. K is a user defined value. It's depended upon user that how many clusters are needed for data processing. The mainly used algorithm is k means clustering algorithm.

Bayesian Clustering technique: In this technique, initialize all data elements into individual cluster with probability arrays. Compute the probability array for each cluster. Merge cluster based on highest probability with that particular cluster array. After that merge all those clusters and have to assign it as the parent of two child cluster. Have to terminate algorithm up to array will not over.

1.2.2] Association: Association is method of discovering inserting relation between variables in large database. It is intended to identify strong rules discovered in database. Association rule inference algorithm is more symbolic in nature, going over sets of items in increasing amity.

1.3 Reinforcement machine learning: Reinforcement learning is comes from supervised learning. But we don't know the output as we know in supervised learning. In reinforcement techniques we only know that to produce output which best actions we have provide. This process also is known as rewarding process. This concept is depends upon behavior of the model. Elements of reinforment learning are 1) Agent 2) Environment 3) Reward 4) state 5) action. We will take example of self driving car to understand concepts of reinforcement learning. Car system is nothing but agent. Road, traffic, signals all these things known as environment. State is like traffic signal means if it will be red we have to stop. And action is after showing green signal car have to run. If our self driving car will be run on green signal that is known as positive reward and on the other hand if it will not run on green signal it may be its negative reward. The main goal of the agent to maximize the expected cumulative reward. Reinforcement learning refers the goal-oriented algorithms. Reinforcement learning solves the difficult problem of correlating immediate actions with delayed return they produce.

2. Model selection: Model selection is depends upon model complexity. Like that over fitting, under fitting, generalization error, validation as procedure for model selection. Suppose we have training data and one machine learning algorithm; by applying that algorithm you get some prediction lets gives name to it as predicton1. The accuracy of applied algorithm is suppose it is 0.24. And as per real word data it provides the accuracy is 0.25. so as compare to these two accuracy like existing algorithm accuracy and we calculated accuracy. Our accuracy is less than existing accuracy. So we have to apply with different machine learning algorithm. Check the accuracy of second prediction as prediction2. And suppose accuracy is 0.95%. Compared with real word data and its accuracy. Our algorithm is providing good accuracy. First prediction is known as under fitting and second prediction is known as over fitting. Those models are selected which provides good accuracy. Models are selected by applying number of machine learning algorithm on train and test set like. Cross validation, k- fold method, train test split method.

3. Prediction Model: The goal of predictive model is that the data we have what will be the output in future. Like that predicting that which type of restaurant will be like more. Prediction model have one of the important part is recommendation engine. That from that recommendation engine we can analyze that which type of food, lighting, music, drinks etc likes to the customer. As per search engine we have to predict our data. We will start clarifying through one example as data analytics; we have the database with more than thousand records. We will consider that data as a dataset. Dataset contains with historical data. Means that those values which we know are already known to particular dataset. Present dataset is used for modern or future data processing. Suppose we have one of the shopping websites. It has lots of products for men, women, kids, home appliances likewise. We have to know that in which product the customer is more interested. Suppose we have new customer handling our websites. How to predict that in which product that particular person is interested. It may be depends upon gender of the person. Age of the person, last history of purchase order. From these things we can predict the values. Many people are referring trending products. All theses information stored into dataset and has to predict by applying machine learning algorithm.

4. Feature selection method: Feature selection method is used to create accurate predictive model. Feature selection is used to remove unnecessary, irrelevant and redundant data from used dataset. Feature selection is used for the purpose of improving prediction performance; provide faster and more cost effective predictors. There are mainly three methods of feature selection algorithms. First one is filter method. In this method, the statically method is applied on given dataset and measure the scoring of each feature. Those who have less scoring will be ignored but it is totally depends upon the algorithm which used in feature engineering. Some of the example of filter method is Chi squared test, information gain as well as correlation coefficient scores. Second method is Wrapper method. This method is used for selection of set of features where checked different combinations are considered and compared with those combinations. From that who has better accuracy will be considered as final wrapper model. Recursive feature elimination algorithm is used in wrapper method. Last model for feature selection is embedded method. Embedded model checks that which feature provide good accuracy at the time of model building. Some of the tools are used for feature selection like weka tool, scikit-learn R packages.

Following are some types of feature selection: 1] Train test split method 2] cross validation method

Train test split method: As per our knowledge we know that training set have known output. And test dataset is used for to test our model's prediction on given test dataset. In train test split model data is divided into two groups like that 70% of data is used as the train data and remaining 30% of data is used as the test dataset. Train test dataset is decided by applying above mentioned three methods

Validation method: It is little bit same as train test split but it creates number of subsets. Subsets are nothing but the test set. Suppose we have dataset with recorded of five thousands. The cross validation method is worked as dataset may be divided into more than two groups. As is considered I have dataset with five thousands record. I will divide into five parts. Five values is considered as k value. From this dataset k-1 part of dataset is used as test dataset and other one remaining is used as train dataset. As per rules it checks the accuracy at every test dataset. This process may be continues up to every single partition is goes for training.

Discussion

1. Selection and Application of machine learning algorithm in production quality.

In this paper, represented that a tangible use case in which ML algorithm is applied for prediction. It predicts the quality of products in process chain. In a process chain consisting of six processes. It should predict after completion of each individual process whether the product would be off-spec in the following process. For implementation decision making tool (DMT) is used.

2. Predicting future hourly residential electrical consumption- A machine learning algorithm

In this paper, report on evolution of the seven different machine learning algorithms is applied on new residential dataset which contains with sensor measurement collected every five minutes. Least squared SVM perform best fit as compare to the entire seven algorithms.

3. Supervised Machine Learning: classification technique.

In this paper describes various supervised machine learning algorithm. The goal of supervised learning is to build a concise model of the distribution of class labels in terms of predictor features. The resulting classifier is then used to assign class labels to the testing instances where the values of the predictor features are known, but the value of the class label is unknown.

4. C5.0 machine learning algorithm

In this paper, implemented the C5.0 algorithm. This algorithm is based on decision tree. C5.0 is implemented new techniques. Like boosting method, in this method several decision trees are generated for the purpose of improving the prediction. New attributes are added in this algorithm like that time stamp, order discrete values, dates etc. This algorithm supports the sampling and cross validation.

5. Comparison of five machine learning algorithm for IP traffic flow classification.

In this paper we recognize that real-time traffic classifiers will operate under constraints, which limit the number and type of features that can be calculated. On this basis we define 22 flows
Features that are simple to compute and are well understood within the networking community. We evaluate the classification accuracy and computational performance of C4.5, Byes Network, Naïve Byes and Naïve Byes Tree algorithms using the 22 features and with two additional reduced feature sets.
6. AdaBoost and random forest as interpolating classifier
In this paper, applied both AdaBoost and random forest algorithm for similar task and check the accuracy of algorithm. Both algorithms achieve same prediction accuracy. But random forest does not conceived as direct optimization procedure. AdaBoost algorithm used static optimization procedure at every stage of algorithm.

7. Predicting review rating for product market.
In this paper used both approaches for comparing the results of the both the system first three columns are non hadoop tabs and same operations are performed using hadoop, map reduce component from terminal, to check the difference in execution. For visualizing the output of sentiment analysis in form of pie chart for the particular product which also display total reviews and rating. A graph chart for comparing the review rating and last tab shows the review ratings are calculated.

8. Machine learning based object identification system
In this set used predefined training and testing data set which is used to predict various types of object. In this paper implemented classification type algorithm that is CNN algorithm. By applying this algorithm the accuracy is 95.5% gives better classification accuracy.

3. CONCLUSIONS

Thus we have studied that machine learning concepts, model selection method. And its type with some example. Feature selection is consist of number of machine learning algorithm like that SVM, decision tree, KNN, adaBoost, Random forest, linear regression, logistic regression. The future scope of this study is applying these algorithms on ISCX2017 dataset to predict the attacks. We have to check that which algorithm provides good accuracy.

REFERENCES

[1] Muhammad Fahad Umer, Muhmmad sher, Yaxin Bi “Flow-based intrusion detection: Techniques and challenges”, June 2017.

[2] Ragini Avhad, “Machine learning”, Feb 2019.

[3] Bharath P, Saravanan M, Aravindhan K, “Literature Survey on smart vehicle Accident prediction using Machine learning algorithm ”, Feb 2019.

[4] Dhivya R, Dharshana R, Divya V, “security Attack Detection in cloud using machine learning algorithms ”, Feb 2019.

[5] Jonathan Krub, Maik Frye, Gustav Toedoro Dohler Beck, Robert H. Schmitt “Selection and Application of machine learning algorithm in production quality.”, Jan 2019.

[6] Richard E. Edwards, Johua New, Lynne E. Parker “Predicting future hourly residential electrical consumption- A machine learning algorithm”, March 2012.

[7] Abraham J. Wyner, Matthew Olson, Justin Bleich “Explaining the success of AdaBoost and random forest as interploting classifier”, Feb 2017.

[8] D. Vetriselvi D. Monish, M. Monish Varshini “Predicting review rating for product marketing.”, March 2019.

[9] K.Rajendra Prasad, P. Chandana Sravani, P.S.N. Mounika, N. Navya, M. Shyamala “Machine learning based object identification system”, March 2019.