A Simple Approach for Selecting the Best Machine Learning Algorithm

Achshah R M, Dr.T.Ruban Deva Prakash

Abstract—. Machine learning algorithm is the soul of artificial intelligence. There are various machine learning algorithms, programmers choose the algorithm that is tailor made for their problem based on performance, pros and cons. This paper discuss the pros and cons of the commonly used ML algorithms: linear regression, logistic regression, k-nearest neighbor, decision tree, random forest, naive bayes, artificial neural network, convolution neural network, support vector machine and XG-Boost. It compares the above supervised learning algorithms explaining the fundamental concepts, thus, making it easy for beginners to understand and choose the right algorithm for their problem. Beginners find it difficult to select the right ML algorithm for their application. However, a simple approach to select the best suited algorithm by analyzing their training data is the aim of this paper. The performance of each algorithm with different training dataset is explored. Based on linearity, outliers, noise, normality of distribution, missing values, dimensionality, speed and preprocessing requirement of training data, the efficient algorithm is selected. It is important that selected algorithm has high accuracy and precision. It is a tremendous and time-consuming work to train each algorithm with the dataset and then choose the one with the best accuracy score. Hence, it is tranquil if one can compare and choose, thus saving time, using the proposed approach.

Index Terms— algorithm selection, artificial neural network, convolution neural network, decision tree, k-nearest neighbor, linear regression, logistic regression, machine learning, naive bayes, random forest, support vector machine, XG-Boost.

1 INTRODUCTION

MAchine Learning (ML) algorithms have various applications in medical, agriculture, banking, business and education Machine learning algorithms can be categorized into different types: supervised learning, unsupervised learning, semi-supervised learning, reinforcement learning and self-supervised learning.

In this paper we are focused on supervised machine learning algorithms. Supervised learning algorithms learn from labelled historical data. Labelled data is the data that has input variables and target variable. The algorithm learns the relationship between the input variable and target variable. Once the algorithm is trained, using past data, it can be deployed for prediction. Supervised learning is the most popularly used method and it can be evaluated and fine-tuned to attain the accurate predictions easily. Supervised learning algorithms are further divided into regression and classification problems. Ayon Dey [1] surveyed all the types of ML algorithms and elaborated in detail. Batta Mahesh [2] explains the various machine learning algorithms with pseudo code and applications. Gyana Ranjan Patra and et al [3] described the advantages and drawbacks of Artificial Neural Network (ANN), Back Propagation Algorithm, Bayesian Learning, Decision Tree and Support Vector Machine (SVM) in application point of view to develop ideal model for Boltzmann machines.

Umer Ahmed Butt and et al [4] reviewed different ML algorithms for overcoming the cloud security issues using supervised, unsupervised, semi-supervised, and reinforcement learning. Finally, compared its performances for this particular application. Darpan Pandey and et al [5] discussed the various algorithms including the preprocessing, feature extraction, training, testing, hyper tuning methods and optimization. S. B. Kotsiantis [6] analyzed classifiers, data mining techniques, and intelligent data analysis. Vineet Chaturvedi and Walter T. de Vries [7] reviewed the ML algorithms used for the application of urban land use planning. Artificial intelligence and machine learning algorithms are used in all applications but selection of ML algorithm for a specific application is challenge for beginners. This paper explains how to find the best algorithm for specific application. It also explains the working of the algorithms in simple terms with its pros and cons. Comparison of the features of various algorithms are also highlighted.

2 REGRESSION PROBLEM

Regression helps to find the correlation between the input variables and target variable when the target variable is continuous or real. For example the target variable is temperature, price, age etc.. The regression actually fits a regression line in such a way that the vertical distance between the regression line and the target dataset is minimum. This vertical distance between the target dataset and regression line is called error. In fig.-1, the dots denote the target data set and the line drawn is regression line. The goal of regression is minimizing the gap between the target data set and regression line.

List of regression algorithms in Supervised Machine Learning.

- Linear Regression.
- Ridge Regression.
- Lasso Regression.
- Neural Network Regression.
- Decision Tree Regression.
- Random Forest.

- *Achshah R M is currently pursuing bachelor's degree program in artificial intelligence and data science in SRM Eswari College of Engineering, India, PH-6238127267. E-mail: samuelachsha108@gmail.com*
- *Dr.T.Ruban Deva Prakash is a researcher in application of AI in power quality and currently serving as Principal in Heera College of Engineering and Technology, India, PH-9446595197. E-mail: hcetprincipal76@gmail.com*

TABLE – 1
ALGORITHM SELECTION TABLE FOR REGRESSION

	Linear	Ridge	Lasso	Neural Network	Decision Tree	Random Forest	KNN	Support Vector
Assumes linear relationship	Yes	Yes	Yes	No	No	No	No	No
Sensitive to outliers	Yes	Yes	No*	Yes	No	No	Yes	No
Sensitive to noise	Yes	No	No	No*	No*	No	Yes	Yes
Solves overfitting	No	Yes	Yes	No	No	Yes	Yes	Yes*
Assumes normal distribution	Yes	No	No	No	No	No	No	No
Speed	Fast	Fast	Fast	Slow	Fast	Slow	Fast	Fast*
Data pre-processing required	Yes	Yes	Yes	No*	No	No*	Yes	No*
Sensitive to missing values	Yes	Yes	Yes	Yes*	No	No	Yes	Yes
Size of dataset	Big	Big	Big	Big*	Average	Average	Small	Small
Works with higher dimensional data	No	Yes	Yes	No	Yes	Yes	No	Yes

*Accuracy is slightly affected or only if it is less than certain values etc.

- K Nearest Neighbors Regression.
- Support Vector Regression (SVR).

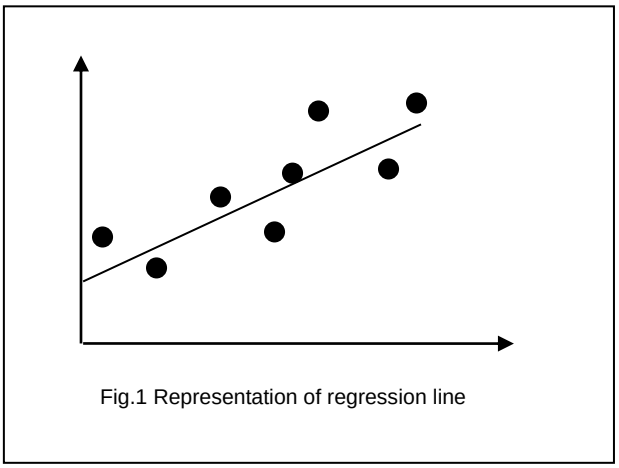

Fig.1 Representation of regression line

The comparison for selection of the above algorithms is given in Table-1

2.1 Linear Regression

Linear regression is the oldest, simple and popularly used supervised machine learning algorithm. Many problems can be transformed into linear regression problems and it handles overfitting pretty well using dimensionally reduction techniques, regularisation, and cross-validation hence it is widely used. One advantage is the extrapolation beyond a specific data set which makes it easy to predict. Linear regression algorithm works best when:

- there is linear (or straight line) relationship between the input variables and target variable as it assumes linear relationship between them,
- the outliers have been removed because it is very sensitive to anomalies in data,
- the noise has been removed as, otherwise it will start modelling the noise instead of the relationship between the variables,
- the number of samples are greater the number of parameters,
- the collinearity has been removed because it will overfit if the input variables are highly correlated with each other and,
- the input and target variables have Gaussian distribution as the predictions will be more reliable.

We use linear regression algorithm for the following:

- Sales Forecasting
- Risk Analysis
- House Price Prediction and other Factors

- Stock Price Prediction
- Analysing Trends
- Real Estate Prediction
- Salary Prediction

2.2 Ridge Regression

Ridge regression is just a better version of linear regression where a small amount of bias is introduced to get better long term predictions. It is actually a model tuning method for the models that suffer from multicollinearity. It is a regularisation method which can reduce the complexity of the model by performing L2_regularization. Ridge regression performs well when:

- there is high collinearity between the input variables,
- number of parameters are greater than or equal the number of samples,
- there is a subset of true coefficients which are small or even zero,
- all the input variables are important as it uses all the input variables and,
- solves the problem of over-fitting.

We use ridge regression algorithm for the following:

- Analysis of Prostate-Specific Antigen
- Agricultural Economics
- Water Resources

2.3 Lasso Regression

Lasso regression is another regularisation method like ridge regression to reduce the complexity of model. It is used to get accurate predictions. It is uses shrinkage. It uses automatic variable selection method and is better than most of the methods. However, the model by lasso regression can be unstable, senseless and not intuitive to interpret. Lasso regression can:

- solve the problem of over-fitting,
- do feature selection that is can reduce the number of input variables,
- be fast in terms of interference and fitting and,
- perform well when number of parameters are greater than the number of samples.

We use lasso regression algorithm for the following:

- Economics and Finance
- Corporate Bankruptcy Prediction
- High Growth Firms Prediction

2.4 Neural Network Regression

Neural networks are popularly used for classification problems but it can be used for regression problems too as it can perform with non linear input variables and target variables. Neural network regression predicts an output variable as a function of the inputs. It has the ability to detect all possible interactions between target variable. Neural network regression performs well when:

- less statistical training is required,
- there are more data points as they are complex,
- the predictions need to be quick (once trained) and,
- there are non linear data with large inputs.

There are some cons to neural network regression, however the pros outweigh the cons. The cons are as follow:

- It has a black box nature which means we can't understand how the variables are connect.
- Does not solve overfitting as the model depends on training dataset.
- Great computational burden as it is expensive and time consuming to train.

We use neural network regression algorithm for the following:

- Sales Forecasting
- Medical Predictions
- Risk Analysis
- Data Validation

2.5 Decision Tree Regression

Decision tree can be used to solve both classification and regression problems. Decision tree regression builds a model in the structure of tree (actually inverted tree) starting from the root node, splitting into decision node and ending at terminal or leaf node. It basically breaks down dataset into smaller and smaller subset while at the same time building as associated decision tree. Though they have overfitting, low accuracy for regression model and are unstable, meaning small changers can create huge impact, they are one of the best and mostly used algorithms as they are easy understand and explain to the client.

Decision tree regression works well even when:

- less or no data preparation or cleaning is done(no preprocessing),
- normalisation or scaling of data is not done and,
- data contains missing or null values.

We use decision tree regression algorithm for the following:

- Business Decision Support
- Healthcare Management
- Pharmacology
- Agriculture

:

2.6 Random Forest Regression

Random forest is also called as random decision forest because it builds multiple decision tree and merges them together to get a more accurate and stable prediction. Is is an ensemble learning for both classification and regression problems. It is complex, no Interpretability and time consuming hence not suitable for real time predictions. Random forest regression can:

- solve the problem of over-fitting,
- be extremely flexible and have very high accuracy,
- good prediction from default hyper-parameter and,
- handle missing data.

We use random forest regression algorithm for the following:

- Banking
- Healthcare
- E-commerce
- Stock Market Prediction

2.7 K Nearest Neighbour Regression

KNN can be used for both classification and regression problems. It works based on feature similarity to predict new data that is how closely the query instance resembles the training data. It is widely disposable in real-life scenarios as it is non-parametric. KNN is simple to implement, flexible and handles multi-class cases. K-Nearest Neighbor:

- is faster than other algorithms as it does not require training (this is why it is called a lazy learner),
- does not work well with large dataset as the computation cost is higher,
- does not perform well with high dimensions as it is hard to calculate distance in each dimension,
- needs feature scaling for accurate predictions and,
- is sensitive to noise, outliers and missing values hence they need to be removed.

We use KNN regression algorithm for the following:

- Concept Search
- Recommender Systems
- Recognition Systems

2.8 Support Vector Regression (SVR)

Support vector regression uses same principal as support vector machine. In SVR, the goal is to find the best fit line and the best fit line require is called hyperplane. Support vector regression works well:

- with higher dimensional data,
- with smaller dataset,
- when noise is removed as they are sensitive to noise,
- with both linear and non-linear data and,
- with outliers.

We use SVR algorithm for the following:

- Bioinformatics
- Protein Fold
- Recognition Systems

3 CLASSIFICATION PROBLEM

Classification predicts the category the data belongs to from the dataset. Classification works only when the target variable is discrete (or categorical). For example the target variable is shoe size, number of children, types of flower etc.

List of classification algorithms in Supervised Machine Learning are:

- Logistic Regression.
- K Nearest Neighbors (KNN).
- Support Vector Machine (SVM).
- Naive Bayes.
- Decision Tree.
- Random Forest.
- XG Boost.
- Artificial Neural Network.
- Convolution Neural Network.

The comparison for selection of the above algorithms is given in Table-2

3.1 Logistic Regression

Logistic regression is kind of like linear regression. Though it is called regression it performs classification. The target variable has to be binary, for example, Yes/No or True/False. It is based on a sigmoid function where the output can vary from positive infinity to negative infinity. It is easy to implement, interpret and train the data. It is also very fast in classifying unknown data. Logistic regression:

- performs well there is a linear relationship between the input and target variable,
- works when the number of samples is greater than number of parameters as otherwise it may lead to overfitting,
- works when there is no distribution between input and target variables as it does not assume any distribution,
- is less prone to overfitting and,
- Can give the measure of relevant as well as the direction of the predictor, like positive or negative.

We use logistic regression algorithm for the following:

- Credit Score
- Medicine
- Gaming
- Text Editing
- Hotel Booking

3.2 K-Nearest Neighbors

Like we already discussed in regression problems KNN can be used for both regression and classification problems. We mostly use KNN for classification as KNN classifies the new data point based on the similarity measured from the previously stored data points as shown in fig.2. The new data point, x, in fig.2 is closer to pink hence will be classified as pink. One of the difficulties in KNN is to find the correct K value. KNN classifier has similar advantages as KNN regressor:

- KNN is instance based learner (lazy learner) that is it does not learning anything during the training period. Therefore, there is no training period for KNN.
- New data can be added without affecting the accuracy of the model.

KNN does not perform well with:

- higher dimensional data and,
- large dataset.

We use KNN classifier algorithm for the following:

- Text mining
- Agriculture
- Finance
- Medical
- Facial recognition

- Recommendation systems (Amazon, Hulu, Netflix, etc)

3.3 Support Vector Machines (SVM)

Support vector machine is based on the concept of decision plane, also known as hyperplane, to set decision boundaries. It finds the hyperplane that maximises the gap (or margin) between the support vetors as shown in fig.3. The greater the margin the better the model. The pros of SVM and SVR are same:

- Very good in dealing with higher dimensional data.
- Works well with smaller dataset.
- It is useful for both linear and non-linear separable data.
- Manages outliers.

The cons of SVM are:

- Sensitive to noise.
- Picking the right kernel could be hard.
- Cannot accommodate word embeddings.

We use SVM algorithm for the following:

- Text Categorisation
- Classification of Images

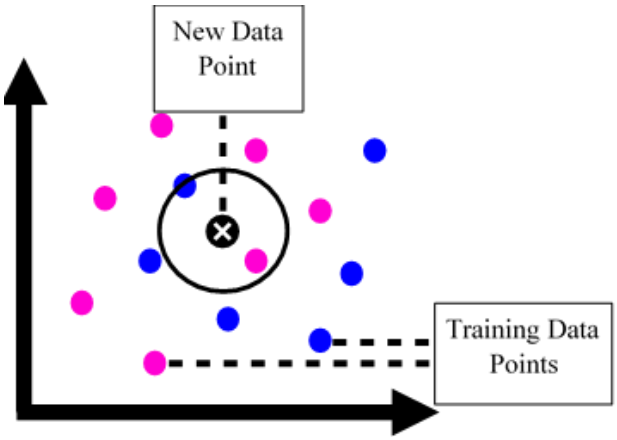

Fig. 2 Representation of KNN

- Face Detection
- Finger Print Recognition
- Bioinformatics

TABLE – 2
ALGORITHM SELECTION TABLE FOR CLASIFICATION

	Logistic	KNN	SVM	Naive Bayes	Decision Tree	Random Forest	XG Boost	ANN	CNN
Sensitive to outliers	Yes	Yes	No	Yes	No	No	Yes	Yes	Yes
Sensitive to noise	Yes	Yes	Yes	No	No*	No	No*	No*	No*
Solves overfitting	No	Yes	Yes*	Yes	No	Yes	Yes	No	No
Speed	Fast	Fast	Fast*	Fast	Fast	Slow	Fast	Slow	Slow
Data preprocessing required	Yes	Yes	No*	Yes	No	No	No*	No	No
Sensitive to missing values	Yes	Yes	Yes	Yes	No	No	No	Yes	Yes
Size of dataset	Big	Small	Small	Big	Average	Average	Average	Big	Big
Works with higher dimensional data	No	No	Yes	Yes	Yes	Yes	Yes	No	No

*Accuracy is slightly affected or only if it is less than certain values etc.

- Handwriting Recognition

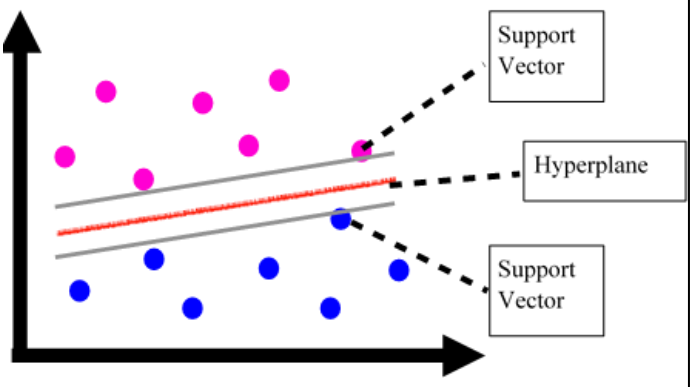

Fig. 3 Representation of SVM

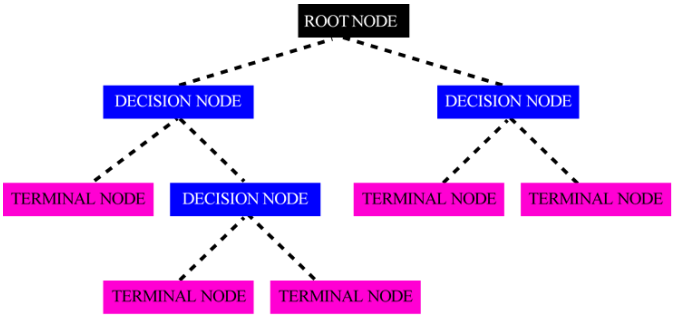

Fig.4 Decision tree

3.4 Naive Bayes

Naive Bayes classifier works based on the concept of Bayes Theorem. The Bayes Theorem says, "the probability that an event A occurs given that another event B has already occurred is equal to the probability that the event B occurs given that A has already occurred multiplied by the probability of occurrence of event A and divided by the probability of occurrence of event B."

$$P(A|B) = (P(B|A).P(A))/P(B)$$

There are three types of Naive Bayes we use: Gaussian NB, Multinomial NB and Bernoulli NB. It is easy and fast to predict the class of test dataset. Naive Bayes performs well:

- in multi class prediction,
- compared to other models like logistic regression (when assumptions of independence holds),
- with less training datasets and,
- with categorical input variable rather than numerical.

The demerits of Naive Bayes are:

- It assumes all the input variables are independent of each other.
- Zero frequency can happen.
- It is known as bad estimator.

We use Naive Bayes algorithm for the following:

- Real Time Predictions
- Multi Class Predictions
- Text Classification
- Spam Filtering
- Sentiment Analysis
- Recommendation Systems

3.5 Decision Tree

As we already discussed in decision tree regression we can use decision tree for classification problems as well. We know that decision tree builds a model in the structure of a tree as in fig.4.

The benefits of decision tree classifier is similar to decision tree regression:

- It is easy and simple to interpret and visualise as it looks like if-else statement.
- No feature scaling, like standardisation and normalisation, is required.
- Can handle non-linear parameters efficiently.
- Can handle missing values and outliers.
- Less training period required compared to other models like random forest as it builds only one tree.

The drawbacks of decision tree classifier are:

- Generally leads to overfitting.
- It is unstable.
- Sensitive to noise.
- Not suitable for large datasets.

We use decision tree algorithm for the following:

- Engineering
- Civil Planning
- Law
- Assessing Prospective Growth in Business

3.6 Random Forest

Random forest is an ensemble learner, based on bagging algorithm, that is it builds multiple models. Random forest classifier operates by constructing a multitude of decision trees at the training time and outputting the class that is the mode of the classes as shown in fig.5. Random Forest was invented to fix overfitting in decision tree.

The advantages of random forest classifier algorithm are:

- Reduces overfitting.
- No feature scaling required.
- Can handle missing values and outliers.
- Handles non-linear parameters.
- Provides high accuracy.
- It is very stable.
- Not sensitive to noise.

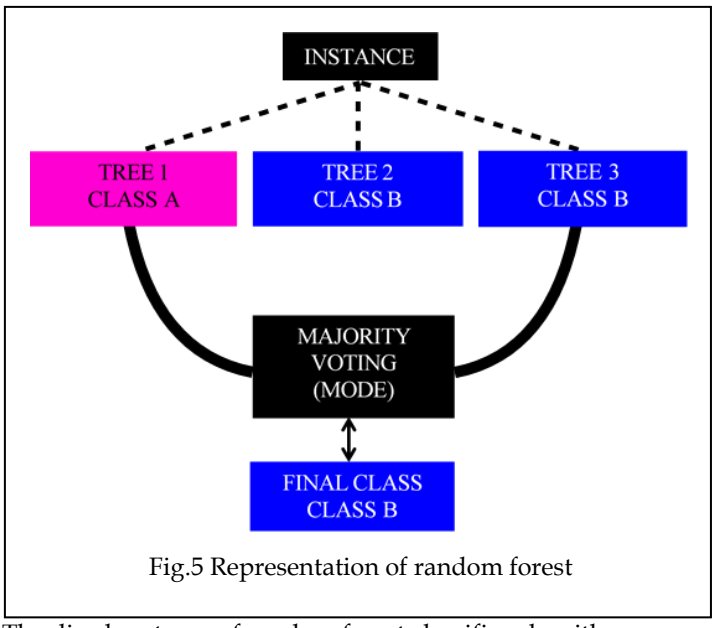

Fig.5 Representation of random forest

The disadvantages of random forest classifier algorithm are:

- High computational power required (and resources).
- Time consuming (longer training period).
- Interpretation is hard (complex).

We use random forest algorithm for the following:

- Credit Card Fraud Detection
- Healthcare and Medicine
- Stock Market Prediction
- Product Recommendation (E-commerce)

3.7 XG Boost

XG Boost is an implementation of gradient boosted decision tree designed for speed and performance. Boosting is an ensemble method like bagging so it combines the predictions of several models (decision trees) into one. XG boost is highly flexible. XG boost performs:

- with missing values,
- with small datasets,
- with overfitting as it has built-in L1 and L2 regularisation,
- faster as it uses multiple CPU cores (parallel processing).
- at optimum as it allows user to perform cross validation and,
- effective tree pruning.

XG boost does not have a lot of disadvantages:

- Sensitive to outliers
- Lack scalability

We use XG boost algorithm for the following:

- Banking Sector
- Medical
- Stock Market
- Business

3.8 Artificial Neural Network

Artificial neural networks are, also called neural networks, inspired from biological neural networks. It is a simulation of biological brain. It takes input signals and does some non-linear functions and sums them to give the output. ANN has three layers: Input layer, Hidden layer and Output layer. In fig.6, the X (X1 and X2) is input signal, w is the wight, B (B1, B2, and B3) is the bias and Y is the output. The following are the advantages of ANN:

- Informations are stored on the entire network hence missing information in one place does not affect the model.
- It can handle missing information but the performance depends on the importance of the missing information.
- They have fault tolerance.
- It has parallel processing ability.
- It has gradual corruption that is the network does not corrode immediately.
- They have distributed memory.

ANN has various disadvantages like:

- Hardware required depends on the structure of the model.
- Reduced trust as it does show how or why the given solution.
- There is no specific rule for determining the structure of ANN.
- The duration of the network is unknown.

We use ANN algorithm for the following:

- Handwriting Recognition
- Traveling Salesman Problem
- Image Compression
- Stock Exchange Prediction

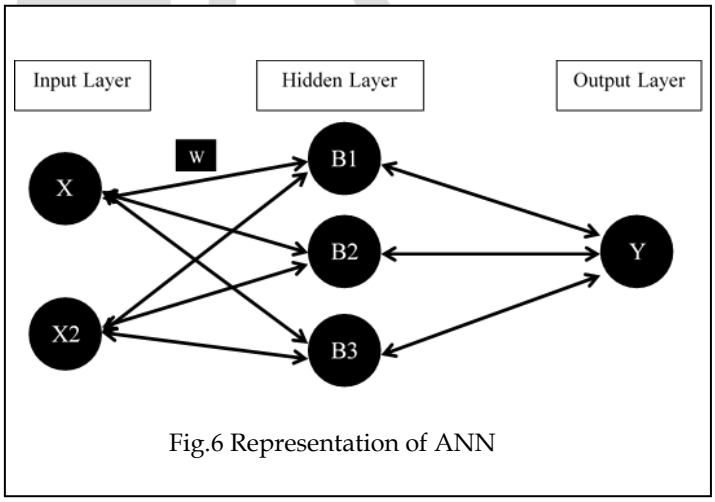

Fig.6 Representation of ANN

3.9 Convolution Neural Network

CNN is similar to ANN but it is mostly used for image recognition. Like ANN they are made of neurone with learnable weights and biases. The CNN contains four layers as demonstrated in the fig.7:

- Convolution layer: The data or image is convolved using filters or kernels.

- Activation layer: The ReLu, Rectified Linear Unit, is applied.
- Pooling layer: The downsampling of features is done.
- Fully connected layer: Flattening is done in this layer. This is the last layer where all the features are combined together to create a model

The pros of using CNN:

- They have high accuracy when it comes to image recognition.
- They can automatically detect the important features without the help of a human.
- They can share weights, happens in a particular layer (filters).

The cons of using CNN:

- They do not encode the position and orientation of the object.
- Large training data is required by CNN.
- Lack of ability to be spatially invariant to the input data.

We use CNN algorithm for the following:

- Search Engines.
- Recommender System
- Social Media
- Face Recognition
- Medical Imaging

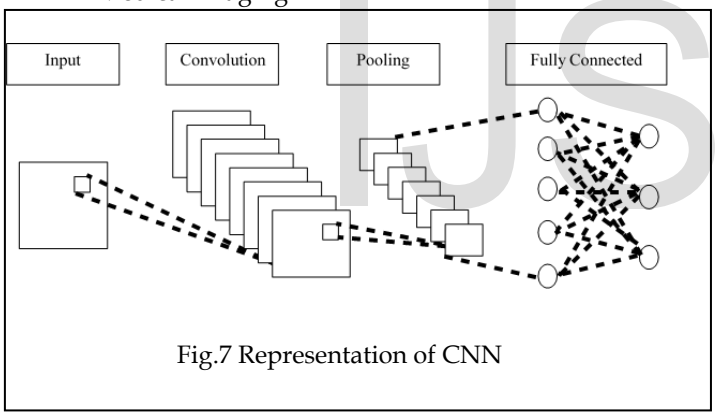

Fig.7 Representation of CNN

4 Procedure for Selecting Best Algorithm

A. Find the correlation to check linearity.
B. Plot a box plot to check for outliers.
C. Find the CV score to check for noise.
D. Find the bias and variance to check for overfitting or under fitting.
E. Find skewness and kurtosis to test normality.
F. Decide on the speed of training based on application, especially for real-time.
G. Check for redundancy, missing values and duplicates.
H. Note the dimension and size of the training data.
I. Using the above values select the best two algorithms by comparing with table 1/ table 2.
J. Perform preprocessing if it is required for the selected algorithm as per selection table.
K. Train both algorithms using the training dataset and select the one with higher accuracy score.

5 Conclusion

Machine learning is a field of computer science that gives computers the ability to learn on their own, through experience and the use of data, without being explicitly programmed. It is an application or an area of artificial intelligence. Now every one are using AI knowingly or unknowingly while using social media, watching Netflix, online shopping etc. Machine learning algorithm is the core of AI and selection of ML algorithm for specific application is the skill of programmer. Most programmers use trial and error approach to find the best suited efficient algorithm which is time consuming. This paper proposed a simple approach for finding the best algorithm by analyzing the training data set and comparing using the proposed approach.

References

[1] Ayon Dey, "Machine Learning Algorithms: A Review", International Journal of Computer Science and Information Technologies, vol. 7, no. 3, pp. 1174-1179, 2016.
[2] Batta Mahesh, "Machine Learning Algorithms - A Review", International Journal of Science and Research, vol. 9, no. 1, pp.1151-1157, January 2020.
[3] Gyana Ranjan Patra, Sarada Prasanna Pati, Mihir Narayan Mohanty, "A Review Paper on Machine Learning Algorithms", International Journal of Advanced Research in Engineering and Technology, vol. 11, no. 12, pp. 197-203, December 2020.
[4] Umer Ahmed Butt, Muhammad Mehmood, Syed Bilal Hussain Shah, Rashid Amin, M. Waqas Shaukat, Syed Mohsan Raza, Doug Young Suh, and Md. Jalil Piran, "A Review of Machine Learning Algorithms for Cloud Computing Security", Electronics by MDPI, The Polish Society of Applied Electromagnetics, vol. 9, no. 1, pp. 1-25, 2020.
[5] Darpan Pandey, Kamal Niwaria, Bharti Chourasia, "Machine Learning Algorithms: A Review", International Research Journal of Engineering and Technology, vol. 6, no. 2, pp. 916-922, February 2019.
[6] S. B. Kotsianti, "Supervised Machine Learning: A Review of Classification Technique", Imformatica, vol. 31, no. 1, pp. 249-268, 2007.
[7] Vineet Chaturvedi and Walter T. de Vries, "Machine Learning Algorithms for Urban Land Use Planning: A Review", Urban Science by MDPI, vol. 5, no. 1, pp. 1-16, September 2021.