ARTIFICIAL INTELLIGENCE AND ITS TECHNIQUES

Bhavana P

Christ University

ABSTRACT:

Life without computer in today's world is unthinkable. The entire domain in today's world including day to day activities is computerized using developing technologies. Along with these developments, the investment and research in self-learning machines for decision making, predicting, planning, analysing, trade and logistics has increased drastically. In future, automated intelligent machines can replace the human activities or improvise the skills of human being. Artificial intelligence is the intelligence exhibited by machines in accordance to algorithm instructed to them and analysing about data given to them. The popularity of artificial intelligence in past two decades has increased enormously. This paper will give clear understanding about the tools or various algorithm used in artificial intelligence. This paper will also explore about the pros and cons of each algorithm and explains in detail about the parameters considered.

Key words: artificial intelligence, human, machines, algorithm

I. INTRODUCTION:

It is said that artificial intelligence is playing a huge role in development of technology in todays world. Intelligent refers to acquiring knowledge or gathering information, in the same way artificial intelligence is the process of inhibiting knowledge to machines using different software's which can help in decision making, predicting, planning, analysing, trade and logistics, and communicate, predict, manipulate by learning from gathering large quantity of data. Artificial intelligence helps machine to work smarter and in more applicable way. The different ways of approaching it can be by using statistical method, artificial neural networks and probability.

II. SURVEY /RESEARCH WORK:

A) K NEAREST NEIGHBOURING CLASSIFICATION (KNN):

K nearest neighbouring classification is one of the basic and effectively used algorithms in artificial intelligence. KNN is also known as lazy learning algorithm. Its main purpose is to analyse the database given, in which the data points are classified into several groups, and to predict the group of newly introduced sample points. This is very useful for data which doesn't have any prior knowledge about the distribution. It is one of the unsupervised algorithms.

PARAMETER SELECTION:

The popularly used parameter in KNN algorithm is the number of neighbouring points i.e. value of k. A small value of k means that noise will have a higher influence on the result. A large value makes it computationally expensive [1]. Simpler way of approaching the value of k is by using thumb rule, k=n^ (1/2) where n is the total number of data points. The other method is k—fold cross validation (KVFC), elbow method.

Algorithm [2] [3]:

Step 1: Start

Step 2: Obtain (known) training dataset and unknown test samples.

Step 3: Define the data set for training dataset.

Step 4: Calculate the distance between training dataset and unknown sample given. Check for all unknown data.

Step 5: Categorize the distances and first k distance and corresponding classes.

Step 6: The test pattern is announced to be of class X if no of distances matching with class X is high.

Step 7: if any unknown test sample remains repeated the step from 2 – 5.

Step 8: stop.

PROS:

- There is no pre assumed theoretical data – hence it is useful for non-linear data
- Simple and basic – to understand and interpret.
- Versatile – used for both classification and regression

CONS:

- Memory consumption is high.
- Stores all data including training data.
- Sensitive to irrelevant features and the scale of data.
- Low – efficiency
- Dependency on good selection of k

B) THE RANDOM FOREST OR RANDOM DECISION FORESTS ALGORITHM:

Random forest constructs multiple trees and combines them to form more accurate and firm estimation. These trees are formed on arbitrary feature basis. The first algorithm for random decision forests was created by Tin Kam [4] Housing the random subspace method in 1995. It is one of the supervised algorithms.

PARAMETER SELECTION [5]:

Feature which make predictions of the model better: (this improves performance of the model and decreases the speed)

i. Max feature:

These are the maximum features that random forest is allowed to take in separate tree. There is multiple option for this feature

ii. Auto/none:

In this option it simply assumes all the features which make sense in every tree. There is no constraint.

iii. Sqrt:

In this option square root of total number of features in one single run is considered.

iv. .2:

It considers 20% of total number of features.

Feature that make model training easier: (this impacts the model training speed)

i. N jobs:

It indicates the engine about the usage the processor it is allowed to use

For example:

For (-1) no restriction on the number of processors.

For (1) it uses only one processor.

ii. **Random state**:

It is the parameter which makes solution to replicate.

iii. Oob score:

It is one of the cross-validation methods. It tags every observation used in different trees and finds the maximum vote.

Algorithm [6] [7]:

Step 1: Start

Step 2: Obtain and read source sample

Step 3: Subsample the given source sample into different groups based on some random characteristics that makes some sense.

Step 4: Construct the data of subsample into decision tree.

Step 5: The predicted results fall into leaves.

Step 6: Predicted results from all constructed trees are gathered and averaged.

Step 7: From the average which ever is the nearest matching the final prediction falls into that particular category.

Step 8: Stop.

PROS:

- The process of averaging or decision tree help to overcome problem of overfitting.
- Less variance compared to single decision tree which means that there is more accuracy even for large range of data items.
- It is flexible and more accurate.
- There is no necessity for scaling down the data
- It is best suited when large proportion of data is missing to maintain the accuracy

CONS:

- It is much harder and time consuming to construct than decision tree hence it is very complex.
- It requires computational resources and also less intuitive when there is large collection of data it is hard to have intuitive grasp of relationship existing input.
- It is very time consuming in predicting process when compared to others.

C) Support Vector Machine (S.V.M):

Supervised vector machine is one of the supervised machine learning algorithm.it can be used for both classifier and regression. Support vectors are the co-ordinates of individual observations. SVM is boundary which segregates two classes (known as hyper plane). The special feature of this algorithm is to ignore outliers and find hyper plane that has maximum region.

PARAMETER [8]:

There are different parameters for this algorithm. Here the three major parameter is considered.

Kernel: There is various option to use this parameter they are as listed below:

Linear: it is usually used for larger data greater than 1000 because it is more likely that these data are linearly discrete

Gamma: kernel co-efficient for rbf, ply, and sigmoid. When the value of gamma increases it accordingly tries to fit exactly as per training data set

C:it is the penalty parameter of the error term. It also controls the trade-off between even decision boundary and classifying training points.

ALGORITHM [9] [10]:

Step 1: Start

Step 2: Give the historical data.

Step 3: Define SVM data formatting and SVM training process.

Step 4: Approach the v class validation format

Step 5: Give the trained model and new data set and let SVM forecasting process occur.

Step 6: The final forecast is obtained

Step 7: Stop

PROS:

- It works clearly when there is margin of separation.
- It is very operative in high dimensional spaces.
- It is very effective in cases where the number of samples is lesser than number of dimensions.
- It is memory resourceful because it uses subsets of training points in decision function known as support vectors.

CONS:

- It requires larger training time when there is large data set hence it doesn't perform well for larger data.
- It doesn't perform well when data set target sets are overlying.
- This doesn't directly provide probability estimates, these are calculated using an expensive 5 folds cross validation method.

D) Naïve Bayes Algorithm:

It was introduced (though not under the name) into the text retrieved community in early 1960's, and remains a popular method for text categorization. It is of highly scalable. It pre determines the specific characteristic feature in class and it is unrelated any other characteristic features.

PARAMETER:

The one of the most efficiently used parameters is alpha which is known as hyper parameter. The best way to determine this parameter is by using grid search over possible parameter value s, using cross validation method.

ALGORITHM [11] [12]:

Step 1: Start

Step 2: for each trait X, traverse trait list for X at inspected node

Step 3: find the probability using the value of X to fit into the group of class.

Step 4: update the class for X

Step 5: if any value is remining, then continue the step 3 and 4.

Step 6: if any attribute is remaining follow the step from 2 to 5.

Step 7: Stop.

PROS:

- It is fast and easy to predict the class of test data set. It also performs well in multi class prediction process.
- When assumptions of independence are considered, the naïve bayes algorithm holds good comparatively.
- It holds good even when there is less training data.
- It holds good for categorical input

CONS:

- Zero frequency: If categorical variable has a category (in test data set), which was not observed in training data set. then model predicts 0 (zero) probability.
- It is also known as bad predictor so the probability outcomes are not to be taken seriously.
- The other limitations are the assumptions considered about the independent predictors. in real life such set of predictors are impossible to find.

BAYES THEOREM:

$P(A_i/B) = [P(A_i)\, P(B/A_i)] \;/\; [P(A_1)\, P(B/A_1) + P(A_2)\, P(B/A_2) + \ldots\ldots + P(A_n)\, P(B/A_n)]$ [13]

- $P(A_i/B)$ is the probability of $a\,i^{th}$ event of A given B.
- $P(A_i)$ is the probability of i^{th} event occurring.
- $P(B/A_i)$ is the probability of B given in i^{th} event of A.
- $P(A_n)$ is the probability of last event (n^{th}) event occurring.
- $P(B/A_n)$ is the probability of B given in n^{th} event of A.

CONCLUSION:

The field of artificial intelligence has given the ability for non-living things to analyse using various techniques such as algorithm. From past two to three decades there is tremendous development in this field and it continues to play major role in various fields for assisting human in his activities. This paper is based on concept of artificial intelligence, algorithms used in various fields and describes briefly the algorithms and parameter used by the algorithm for tuning and describes the pros and cons of the each algorithm. I conclude that further research in this area could be done as there is very promising and profitable that are obtained by using this algorithm in various fields. We have not yet realized the full potential and capability of artificial intelligence. This has greater impact on human life in the years to come.

References

[1] Imran, "Analytics vidhya," 3 august 2015. [Online]. Available: https://discuss.analyticsvidhya.com/t/how-to-choose-the-value-of-k-in-knn-algorithm/2606/2. [Accessed 16 may 2019].

[2] sankara subbu, ramesh, brief study of classification algorithms in machine learning, 2017.

[3] paul,sudip & sultan,tanin & tahmid,marzana, "automatic electrical home appliance controland securityfor disabled using electroencephalogram based brain-computerinterfacing," in *international conference on advanced information and communication technology(ICAICT)*, Chittagong, 2016.

[4] T. k. Ho, "Random decision forests," in *3rd international conference on document analysisand recognition*, Montreal, 2016.

[5] t. srivastava, "analytics vidhya," 9 june 2015. [Online]. Available: www.analyticsvidhya.com. [Accessed 23 may 2019].

[6] li,ke & yu,nan & li ,pengfei &song, shimin & lei,wu & li, yang& liu, meng, "multi label spacecraft electrical signal classification method based on DBN and random forest," *creative commons attribution 4.0 international* ,

2017.

[7] sammy ongaya, "galaxy data technologies," massive data technologies , 19 6 2018. [Online]. Available: http://galaxydatatech.com. [Accessed 22 may 2019].

[8] "scikit-learn-documents," [Online]. Available: https://scikit-learn.org. [Accessed 23 may 2019].

[9] xydas , erotokritos & marmaras , charalampos &cipcigan , liana &sani hassan, abubakar &jenkins ,nick, "forecasting electric vehicle charging demand using support vector machines," in *universities power engineering conference*, 2013.

[10] "adavanced intelligent computing theories and application with aspects of artificial intelligence," in *4th international conference on intelligent computing ICIC*, shanghai, 2008.

[11] karim , masud & rahman ,mahammad, "decision tree and naive bayes algorithm for classification and generation of actionable knowledge for direct marketing," *journal of software engineering and application*, pp. 169-206, 2013.

[12] sarosa , mohammad & junus , mochammad & hoessny , marina & sari , zamah & fatnuriyah , martin, "classification technique of interveiwer - bot result using naive bayes and phrase reinforcement algorithms," *international journal of emerging technologies in learning*, 2018.

[13] T. Matsuoka, "Baye's theorem and its application to nuclear power plant safety," *international journal of nuclear safety and simulation*, vol. 4, pp. 203-210, 2013.