ISSN(Online): 2319-8753
ISSN (Print) : 2347-6710

International Journal of Innovative Research in Science, Engineering and Technology

(A High Impact Factor, Monthly, Peer Reviewed Journal)

Visit: www.ijirset.com

Vol. 6, Issue 11, November 2017

A Survey on Machine Learning Technique to Predict Number of Bicycles in a Bicycle Rental System

Harish Manghnani, Minal Vanage , Akshata Naik

B. E Student, Department of Computer Engineering, Sinhgad Academy of Engineering, Pune, India

ABSTRACT: The bicycle hailing system has been approved by many countries and is commonly in use like China's RedBike company or Barcelona's Bike hailing system. In this paper we study algorithms to implement on a data set using machine learning to predict number of bicycles at each station in a bicycle rental system. Also we would study to attain the approximate predictive values of number of bicycles at each station. Implementation of algorithms like Naive Bayes will be related to important factors like environment, weather, real time traffic situation and prime timings. On the basis of these factors we will be having training data set and testing dataset which will make our machine learn to predict values with accuracy as high as possible.

KEYWORDS: Machine Learning, Supervised method and overview of unsupervised learning, training dataset and testing dataset for the Neural Network, concepts like gradient descent and backpropogation to improve accuracy.

I. INTRODUCTION

With increase in number of urban traffic PBS (Public Bicycle System) has been implemented by many countries. This provides a solution to problems like last mile or travel small distances effectively faster instead of waiting in traffic congestion for longer period. One of the main objective is to reduce pollution, balance the degrading environment and also contribute to user's fitness regime. One of the best example of PBS is currently working in China Hangzhou where it has more than 3000 bicycles stations and as many as 70000 bicycles in use [1] or PBS of Barcelona called Bicing having nearly 400 stations and 6000 bicycles [3].

Now for instance suppose a bicycle station has 5 bicycles but only 3 users on average daily then the other 2 bicycles are ideal, instead they could be sent to the next station. On the contrary if a particular station has only 3 bicycles but 5 users on daily basis then this degrades the PBS efficiency. This draws us to the conclusion that it is important to know at which station how many bicycles are needed. This is done by predicting number of bicycles per station for owner to attain most efficient working PBS environment.

II. OVERVIEW

A. Requirements-

The implementation includes a python platform where the neural network is designed and implemented. After the formation of the neural network we would be passing the training dataset to train our machine so that our machine learns in real time. The neural network will then be given the testing dataset where at the end in the output it will present a predicted random number and accuracy of the neural network.

B. Design Phase-

In this phase the neural network will be designed for the owner of bike sharing company where he will get an idea of what numbers of bicycles to maintain at each station, as output. The design phase would include various supervised methods like K-means clustering to classify the users [1], or a naïve Bayes algorithm implementation or a simple linear regression, random forest [2] . Main objective will be improving the accuracy of the neural network.

ISSN(Online): 2319-8753
ISSN (Print) : 2347-6710

International Journal of Innovative Research in Science, Engineering and Technology

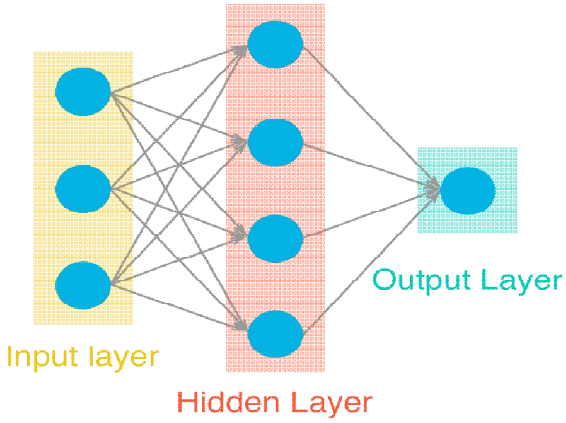
(A High Impact Factor, Monthly, Peer Reviewed Journal)

Visit: www.ijirset.com

Vol. 6, Issue 11, November 2017

C. Implementation phase-

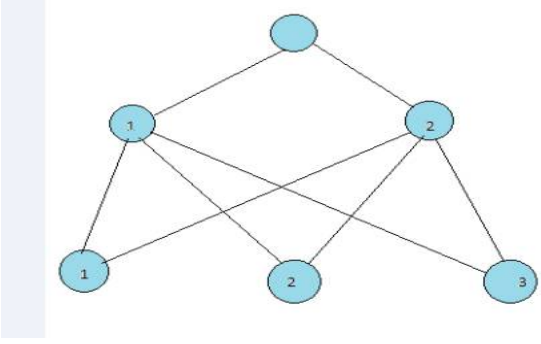
Here the neural network will be added with concepts like gradient descent and back propogation. Gradient descent is a concept where the coding is done in such a way to minimize the error factor. After implementation the primary goal will be to reduce the error and improve accuracy. Here comes the concept of back propogation where the neural network back traces the predicted weights of a particular edge. Basically our neural network will be a less complicated and easy network which has three layers viz. input layer, hidden layer and the output layer. This helps us to understand gradient descent and back propogation easily as there are less perceptrons. Perceptrons are the smallest element of the neural networks connected through edges having weights. In case the error function goes high we can back trace and chose a different path which causes minimum weight, minimum error and maximum accuracy.



III. DATASET

In the dataset we will be introducing two different sets, one which is training dataset which will be training our designed neural network and the other will be testing dataset. We have created two csv files day.csv and hour.csv. these files have important factors that affect the bicycle hailing system like environmental factors temperature, humidity wind speed, weekday, season, year, month etc. These factors are important because for example on weekends people tend to ride bicycle except the rush hours, or for example a pleasant weather encourages a rider to grab a bicycle. Hour.csv helps us to understand the prime hours when people tend to grab a bicycle to reach the nearest subway station or taxi-stand whereas during sunny afternoon people would avoid even getting out of their house and would prefer an evening ride. Training dataset will train the neural network and later when we pass the testing dataset we get appropriate prediction. To improve accuracy and attain reliability we use gradient descent to minimize error. If for example we are using a K-means classification then the clusters will be formed on the basis of classes available for instance 2nd and 4th Saturday off or temperature cluster of 10-20 degrees and 20-30 degrees. Also on basis of weather data can be classified as more users on a non-rainy day as compared to less users on rainy day.

IV. MATH



ISSN(Online): 2319-8753
ISSN (Print) : 2347-6710

International Journal of Innovative Research in Science, Engineering and Technology

(A High Impact Factor, Monthly, Peer Reviewed Journal)

Visit: www.ijirset.com

Vol. 6, Issue 11, November 2017

[$x1$ $x2$ $x3$] * $w11$ $w12$ $w21$ $w22$ $w31$ $w32$

A. Sigmoid (**X**) = $\frac{1}{1+e^{-x}}$

B. Sum of squared errors (SSE)

$$E=\frac{1}{2}\sum \mu(Y^{\mu} - Y^{-\mu})^2$$

C. Gradient Descent

$\Delta w_i = \eta\delta x_i$

$$\delta= (y-\hat{y})\, f(h)$$

$f(h) = \Sigma(w_i x_i)$

Where f(h) is the Activation Function

D. Back Propogation

Here we use the concept of back ward pass
Error= target-output
Output error term= error * output * (1-output)
Hidden_error_term=output_error_term* weight_hidden * hidden_layer_output * (1-hidden_layer_output)

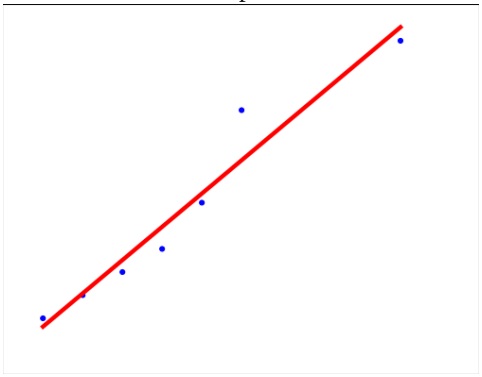
E. Change in Learning Rates

Delta_weight_hidden_to_output = learnrate * output_error_term * hidden-layer-output
Delta_weight_input_to_hidden = learnrate * hidden_error_term * x

V. METHODOLOGY

Machine learning implementation has a vast aspect where we commonly have sub concepts used like supervised and unsupervised learning
Examples of supervised learning are linear regression where the classification is basically done on the basis of a simple equation and we get the result either true or false
Linear Regression = y=mx + c

A simple equation where y is the output or the predicate and x is the independent variable whereas c is just a constant. Linear classification is the easiest classification supervised method used in machine learning.



ISSN(Online): 2319-8753
ISSN (Print) : 2347-6710

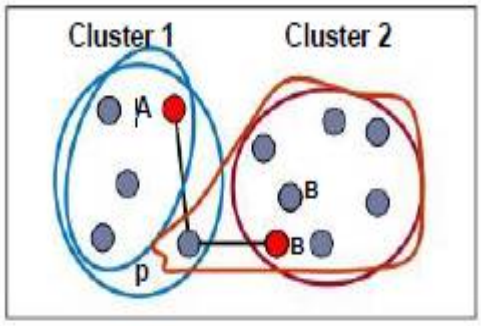
International Journal of Innovative Research in Science, Engineering and Technology

(A High Impact Factor, Monthly, Peer Reviewed Journal)

Visit: www.ijirset.com

Vol. 6, Issue 11, November 2017

Other commonly used supervised methods are K-means clustering algorithm. In this algorithm there are clusters formed. When there is a data entry then the distance of this entry is found with each cluster boundary. Later the smallest distance is evaluated is used as the cluster for that data entry.here in this algorithm the z-score is a important factor to focus on where the z-factor should be in the range -3 to 3 which is considered to a optimal solution. If it is not achieved then it is obvious to choose a different algorithm.

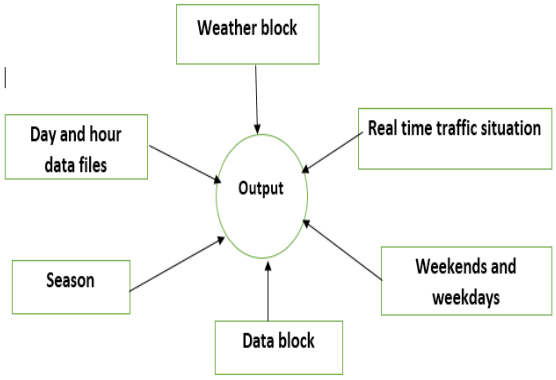


Another popular method used is Random Forest Search where the data entry is searched among existing data where it best suits fits in. it is mostly used supervised method currently in used among the supervised learning methodologies of machine learning and predictions.

Instead we can also use various other forms of classification methods like a Decision Tree in which the classification is based on the data structure as tree or we can use algorithms like extreme gradient boosting where the gradient descent is made to be almost zero so as to reduce the error term to the least and get maximum accuracy

On the other hand we have a mixture of algorithms to improve the output required. For example one of the combination of algorithms include linear regression and Bayesian classification or another example is of support vector machine (SVM) and Bayesian classification. Basic aim of all algoritms is to first classify the data and accordingly predict for the future. Classification is done on the input dataset which is training datatset on which the neural network will be trained and prediction will be done during the testing phase.

VI. ARCHITECTURE



The above architecture clearly shows that the output and the prediction will be basically on the basis of the dataset ie the weather condition, real time traffic, weekends or weekdays, rush hours and also seasons. At the end the

ISSN(Online): 2319-8753
ISSN (Print) : 2347-6710

International Journal of Innovative Research in Science, Engineering and Technology

(A High Impact Factor, Monthly, Peer Reviewed Journal)

Visit: www.ijirset.com

Vol. 6, Issue 11, November 2017

accuracy will be printed where the prediction after training the machine which is a neural network has to be as high as possible.

This is achieved by adding simple details but very effective for a person to decide whether he or she is willing to take a bicycle ride or no. for example people would avoid take a bicycle ride in monsoon when it is raining instead of taking a ride in a pleasant Sunday afternoon where sun shines bright.

VII. SURVEY

Sr.No	ALGORITHM	ACCURACY
1	Random Forest	0.799
2	Decision Tree	0.763
3	Extreme Gradient Boosting	0.749
4	SVM	0.642
5	Linear and Bayesian classification	0.786

The above table clearly shows the accuracy levels of each algorithm used. These accuracy are taken from references and are approximations depending on the dataset and the data entries. The training dataset first trains the neural network and later with the help of above algorithms we apply the testing dataset and predict the number of bicycles needed. The output will be clearly the accuracy of the prediction and our main aim will be getting the highest accuracy with the least error terms. Different algorithms have different ways of determining the mean error term and different ways to minimize them.

VIII. CONCLUSION

The bike rental system is hot system is effectively a topic on which many smart city proposals are working on across the world especially after giant cab hailing service like uber the focus has shifted to a similar concept which not only contributes to the fitness of people riding bicycles but also lessening the effect of urban traffic and pollution. This system reduces the problem of long waiting hours in heavy traffic or also solves problems like last mile problems where people would need 5 minutes of ride instead of 20 minutes of walk. The only factor to be considered throughout the system is the bike availability and security of bikes.

IX. FUTURE SCOPE

The future scope of this project can be implemented in a two wheeler bike rental system in a heavy populated two wheeler user's area like India. The system can could allow users to share a bike travelling in same direction or use the bike to the nearest public transport station.

REFERENCES

[1]. Predicting Occupancy Trends in Barcelona's Bicycle Service Stations Using Open Data Gabriel Martins Dias, Boris Bellalta and Simon Oechsner (IEEE 2015)
[2]. Predicting Public Bicycle Rental Number using Multi-source Data Fei Lin∗, Shihua Wang∗, Jian Jiang∗, Weidi Fan∗, Yong Sun(IEEE 2017)
[3]. Machine Learning or Discrete Choice Models for Car Ownership Demand Estimation and Prediction (IEEE 2017)
[4]. Supply-Demand Prediction for Online Car-hailing Services using Deep Neural Networks Dong Wang , Wei Cao , Jian Li1 Jieping Ye2(IEEE 2017)
[5]. Car Pooling based on Trajectories of Drivers and Requirements of Passengers Fu-Shiung Hsieh (IEEE 2017)
[6]. Characterising Demand and Usage Patterns in a Large Station-based Car Sharing System Chiara Boldrini Raffaele Bruno Marco Conti (IEEE 2016)