VIVA-TECH IJRI
ANNUAL RESEARCH JOURNAL

VIVA-TECH INTERNATIONAL JOURNAL FOR RESEARCH AND INNOVATION
ANNUAL RESEARCH JOURNAL
ISSN(ONLINE): 2581-7280

A Comprehensive review of Conversational Agent and its prediction algorithm

Aditya M. Pujari[1], Rahul M. Dalvi[1], Kaustubh S. Gawde[1], Tatwadarshi P. Nagarhalli[1]

[1](Computer Engineering Department, VIVA Institute of Technology, India)

Abstract: *There is an exponential increase in the use of conversational bots. Conversational bots can be described as a platform that can chat with people using artificial intelligence. The recent advancement has made A.I capable of learning from data and produce an output. This learning of data can be performed by using various machine learning algorithm. Machine learning techniques involves construction of algorithms that can learn for data and can predict the outcome. This paper reviews the efficiency of different machine learning algorithm that are used in conversational bot.*

Keywords *– Machine learning, Random Forest, Linear Regression, K-means, Chatbot.*

1. INTRODUCTION

Artificial Intelligence is also known machine intelligence, is intelligence demonstrated by different machine [14]. Artificial Intelligence is defined as the research of intelligent agents, a device that can learn information from environment and performs action that maximize the chances of successfully achieving its goals. Modern machine capacities are generally classified as artificial appends successfully understanding human speech, competing at the highest level in strategic game, independently operating cars, and intelligent routing in content delivery networks and military simulations. Artificial intelligence research has been divided into subfields that often fail to communicate with each other. These sub-fields are constructed on technical consideration, which includes particular goals, the use of particular tools, or deep philosophical differences [14]. Artificial Intelligence often revolves around the use of algorithms. An algorithm is a set of unambiguous instructions that a mechanical computer can execute. Complex algorithm is basically built on top of other, simpler algorithms. Artificial Intelligence algorithm are capable of learning from data; they can enhance themselves by learning new heuristics, or can themselves write different algorithms. Some of algorithm used Bayesian network, decision trees and nearest-neighbor. This learning of data using different algorithm is known as machine learning.

Machine learning is an interdisciplinary field that uses statistical techniques to give computer systems the ability to learn from data, without being explicitly programmed. Machine learning explores the study and construction of algorithm that can learn for and make prediction on data-such algorithms overcome following strictly static program instructions by making data-driven and make prediction or decisions, through building a model from sample inputs [13]. Machine learning is employed in a range of computing tasks where designing and programming explicit algorithm with good performance is difficult or infeasible. Machine learning is related to computational statistics, which also emphasis on prediction-making past the use of computers. Within the field of data analytic, machine learning is a method used to devise complex models and algorithms that lend themselves to prediction; in commercial use, this is known as predictive analytics [13]. These analytical models let researchers, data scientists, engineers, and analysts to "construct reliable, repeatable decisions and results" and uncover "hidden insights" through learning from historical relationships and trends in the data.

2. CONVERSATIONAL AGENT AND ITS PREDICTION ALGORITHM

2.1 Chatbot for University Related FAQS [18]

Artificial Intelligence conversational agents are becoming popular for web services and systems like scientific, entertainment and commercial systems, and academia. But more effective human-computer interface will take place by querying missing data by the user to provide satisfactory answer. User inquiries are first taken care by AIML check piece to check whether entered inquiry is AIML script or not. AIML is characterized with general inquiries and welcome which is replied by utilizing AIML formats

2.2 Intelligent Chatbot for Easy Web-Analytics Insights[19]

The bot responses were analysed and were classified as wrong, default and correct. The correct responses are the expected domain related answers. More importantly they were the correct answers. Whereas the wrong responses are the bot response to the domain related answers which are wrong. This can be avoided by further refining the bot. The default responses are the bot response to the general or unrelated queries. In this paper, chatbot would enable bot users to just type in the query related to web analytics and will get response immediately. This is to avoid the time consuming task of mastering a web analytics tool. The proposed chatbot is developed using AIML and the data set is the raw analytics data.

2.3 Weather Monitoring using Artificial Intelligence [5]

The paper define weather prediction based on the previous dataset. They intend to develop an Intelligence Weather predicting module since this is become a necessary tool. This tool consider measure such as Maximum Temperature, Minimum Temperature and rainfall for sample period of day on the available dataset and it is done by using machine learning techniques. The prediction and analysis is based on the Linear Regression which predict the next day's weather with good accuracy which is more than 90% is obtain on the dataset. This technique is also gives better performance than the traditional statistical method. In this paper simple Linear Regression is use for weather prediction and it will give better accuracy.

2.4 Rainfall prediction based on 100 years of Meteorological Data [8]

This paper mainly focused on use of the Data Mining techniques for predicting rainfall of an area on basis of some dependent feature like precipitation and wet day frequency. Instead of taking current data they are mainly focusing on the data collected till date from last 100 years' data by the meteorological department. The paper defines Data Mining technique using the Liner Regression Model on the data collected for wet day frequency, precipitation and rainfall. The Liner Regression Model developed has been trained and validate against the actual rainfall of the area, which further was used to predicting the rainfall in coming years.

The paper use Data mining technique with liner regression for predicting the rainfall and it mainly focus on past dataset.

2.5 Rainfall prediction using modified Linear Regression [7]

Rainfall prediction on the historical data is trending in research point of view. The existing model use the data mining technique for predicting the state of atmosphere at a given time of a weather variable like rainfall, cloud conditions, temperature etc. The main problem here in this system is it does not provide an estimate of the predicted rainfall. The proposed system calculates average of the value and understand the state of atmosphere. In this paper it will use mathematical method called linear regression to predict the rainfall in various state of India. The model provides estimate rainfall using different atmospheric attribute like average temperature and cloud cover to predict the rainfall. The main advantages of this model is that the model estimate the rainfall based on the previous correlation between the different parameter.

It uses the modified liner regression to perform the prediction of a rainfall where training and test data are formed from the input data set.

2.6 Regression technique for the prediction of the Stock Price Trend [6]

The paper examines the theory and practice of the regression technique for prediction of stock price by using the transformed data set in ordinal dataset. In this the original pre-transformed data source contain data of heterogeneous data type use for handling of currency value and financial ratio. The data format in currency value and financial ratio is used for computation of stock price. The data source are corporate annual reports which include balance sheet, income statement and cash flow statement. The paper showed that the outcomes of regression technique can be improve for the prediction of stock price.

The outcome of regression technique can be improved when the input data was standardized into a common data type through a customized transformation technique.

2.7 Prediction of Road Traffic Congestion Based on Random Forest [2]

The paper explains how the problem of traffic congestion is solved by using classification algorithm random forest. The city area is divided in sections and prediction is done of the areas possible of having heavy traffic. This is done by considering the environmental conditions such as Climate, Holiday, Road Condition etc. The results show that the traffic prediction model established by using the random forest classification algorithm has a prediction accuracy of 87.5%, and the generalization error is low, and it can be effectively predicted.

The random forest algorithm has the characteristics of high robustness, high performance and high practicability and it will efficiently predict the road traffic congestion.

2.8 Discovery and Prediction of the Unused Land for Construction Based on Random Forest [9]

In the management of land resources, not only solving the existing problems of the land, instead also prediction of the problems of the land and prevention on land misuse are in demand urgently due to the urbanization so that the propose system use Random Forest algorithm for prediction The main part in getting the accurate result of this paper is the formation of the "Decision tree". The most frequently used attribute selection measures in decision tree induction are the Information Gain Ratio criterion and the Gini Index. Random Forest classifier uses the Gini Index as an attribute selection measure. The utilization rate and the period (construction days) between land supply and construction are calculated.

The use of Random Forest for the prediction of unused land is in trend. As the algorithm helps give maximum output accuracy for the required search it is of great use.

2.9 Random forest classification of urban landscape using Landsat archive and ancillary data: Combining seasonal maps with decision level fusion. [1]

In this paper the problem of urbanization is solved by using Random Forest algorithm along with Landsat archive and ancillary data. It proposes a methodology to map the urban areas with multi-seasonal Landsat data. The Random forest classifier and decision level fusion are applied. Shifting importance measure are used to recognise the most important input layers. The annual maps have moderately high (>60%) overall accuracies (OA). The proposed method can be successfully transferred to other years. The methodology was applied to produce a detailed land use land cover map of National Capital Region of India. A second classification involving seasonal maps with decision level fusion based on expert knowledge resulted in an annual composite map with increased number of LULC classes.

The paper gives the general idea about the random forest algorithm of urban landscape.

2.10 Predicting Passenger Flow using Different Influence Factors for Taipei MRT System [10]

According to the statistical data, each day there are over one million passengers taking the MRT in Taipei. In this paper, we will be predicting MRT passenger flow with random forest, by using different factors collected from the Taipei Main station as input for training. In this paper, there are two categories of the factors, which are main factor and support factor. The main factor is the history of passenger flow, which strongly influences the results. As for the support factors is the temporal factor, such as the month and week and the holiday factor, which can slightly increase the accuracy of the prediction.

In this paper, system use only the Taipei main station passenger flow to test the method. Here by taking into consideration it can predict the approximate traffic of the people on occasion's such as public holidays, festivals, week days and week offs etc.

2.11 A Novel Clustering Algorithm Combining Niche Genetic Algorithm with Canopy and K-means [4]

Each chromosome is made up of a sequence of genes coding. The number of genes of a chromosome is randomly chosen where n is the number of data points, which is randomly selected a given data sets. Canopy is usually employed to capture the number of clusters. The canopy method in this paper can automatically capture the number of clusters as the initial population selection and does not need as an input parameter for the number of clusters, which is used to improve multi peak values of canopy. The canopy method to randomly select a number of clusters (K). Chromosomes are selected, which consist of the different number of clusters (K) using canopies.

The paper defines Clustering Algorithm with Canopy which use for making effective cluster.

2.12 An improved K-means Cluster-based Routing Scheme for Wireless Sensor Networks [3]

This paper proposes a cluster-based routing scheme based on an enhanced version of K-means approach. The improved version of K-means generates balanced clusters in the network, which does not overload one cluster-head over the others unlike LEACH where one of the generated clusters may contain a large number of nodes and another contains a small number of members. Besides, the election of the cluster-heads is done in a distributed manner after each period. This paper proposes a cluster-based routing scheme based on an enhanced version of K-means approach.

The improved version of K-means generates balanced clusters in the network, which does not overload one cluster-head over the others unlike LEACH where one of the generated clusters may contain a large number of nodes and another contains a small number of members.

2.13 K-Means Clustering Algorithm Based on Improved Cuckoo Search Algorithm and Its Application [11]

K-means algorithm is a clustering algorithm based on partition. Because of its simplicity and efficiency, it has become one of the most widely used clustering algorithms. However, there are two shortcomings in the original K-means algorithm: firstly, the difference between each clustering result is greatly affected by the initial class centre; secondly, it is easy to fall into the local optimal solution. The original cuckoo algorithm is influenced by step size A and probability of discovery P, and the step size and discovery probability control the accuracy of CS algorithm global and local search, which has great influence on the optimization effect of algorithm. The step size and discovery probability of the CS algorithm are set to a fixed value at initialization and will not change in subsequent iterations. When the step size is set too large, reducing the search accuracy, easy convergence, step length is too small, reducing the search speed, easy to fall into the local optimal.

K-Means algorithm is easy to fall into the local optimum and the Cuckoo search (CS) algorithm is affected by the step size

2.14 An Improved Sampling K-means Clustering Algorithm Based on MapReduce [12]

In the literature present an agglomerative fuzzy K-means clustering algorithm for numerical data, an extension to the standard fuzzy K-means algorithm by introducing a penalty term to the objective function to make the clustering process not sensitive to the initial cluster centres. The MapReduce programming model is mainly composed of two abstract classes, Mapper and Reducer. Mapper processes the cut raw data, Reducer summarizes the intermediate results produced by Mapper and obtains the final result.

The paper extends the K-means clustering process to calculate a weight for each dimension in each cluster and use the weight values to identify the subsets of important dimensions that categorize different clusters.

3. ANALYSIS TABLE

Sr. No.	Technical paper name	Technique Used	Dataset Used	Accuracy
1.	Weather Monitoring Using Artificial Intelligence [5].	Simple Linear Regression	Weather Dataset, Meteorological Department Database.	90%
2.	Rainfall prediction based on 100 years of Meteorological Data [8].	Linear Regression, Data mining, K Fold technique.	Rainfall Dataset, Meteorological Department Database.	NA
3.	Rainfall prediction using modified Linear Regression [7].	Modified Linear Regression.	Rainfall Dataset, Meteorological Department Database.	NA
4.	Regression technique for the prediction of the Stock Price Trend [6].	Regression technique	Companies Dataset; in Bursa, Malaysia.	NA
5.	Prediction of Road Traffic Congestion Based on Random Forest [2].	Random forest.	1124 data from different sections of Shanghai traffic management Department	87.05%

6.	Discovery and Prediction of the Unused Land for Construction Based on Random Forest [9].	Random Forest, Data mining	3 different data sets used	75.73 91.41 84.56
7.	Random forest classification of urban landscape using Landsat archive and ancillary data [1].	Decision level Landsat Multi-season Random forest	Landsat Dataset	>60%
8.	Predicting Passenger Flow using Different Influence Factors for Taipei MRT System [10].	Random forest	Daily passenger data	94.3%
9.	A Novel Clustering Algorithm Combining Niche Genetic Algorithm with Canopy and K-means [4].	Niche genetic algorithm; k-means clustering; Canopy	Iris, Balance, and Breast Cancer UCI database.	87%
10.	An improved K-means Cluster-based Routing Scheme for Wireless Sensor Networks [3].	K-Means.	Energy Consumption.	N.A
11.	K-Means Clustering Algorithm Based on Improved Cuckoo Search Algorithm and Its Application [11].	K-Means; cuckoo search	Iris, Wine, Seeds, Haber man. UCI database	N.A
12.	An Improved Sampling K-means Clustering Algorithm Based on MapReduce [12].	K-means; parallel sampling;	Bag of Words data set in the UCI machine learning library	85%

4. Conclusion

The paper provides a brief survey on various machine learning algorithm which are Random Forest, Linear Regression, and k-means. The survey gives an outcome that larger dataset will provide better result. The survey has shown that numeric prediction, categorical prediction or classification and clustering can be implemented using machine learning algorithm. The study has shown that Random forest gives better results in case of larger dataset whereas K-means is faster clustering algorithm comparing other unsupervised learning algorithm. Simple Linear regression provide better accuracy on different datasets.

References

[1] A. Ghosh, R. Sharma, P.K. Joshi, “Random forest classification of urban landscape using Landsat archive and ancillary data: Combining seasonal maps with decision level fusion”, Applied Geography Journal, 2014, pp. 31-41.

[2] Y. Liu, H. Wu, “Prediction of Road Traffic Congestion Based on Random Forest”, 10th International Symposium on Computational Intelligence and Design, 2017, pp. 361-364.

[3] M. Lehsaini, M.B. Benmahdi, “An improved K-means Cluster-based Routing Scheme for Wireless Sensor Networks”, IEEE, 2018.

[4] H. Zhang, Z. Zhou, "A Novel clustering algorithm combining Niche genetic algorithm with canopy and K-means", International Conference on artificial Intelligence and Big Data, 2018, pp. 26-32.

[5] T.R.V. Anandharajan, G.A. Hariharan, K. K. Vignajeth, R. Jitendiran, "Weather Monitoring Using Artificial Intelligence", International Conference on Computational Intelligence and Networks, 2016.

[6] H. L. Siew, M.J, Nordin, "Regression Techniques for the Prediction of Stock Price Trend", International Conference on Statistics in Science, Business and Engineering (ICSSBE), 2012, pp. 1-5.

[7] S. Prabakaran, P. N. Kumar, P. S. M. Tarun, "Rainfall Prediction Using Modified Linear Regression", ARPN Journal of Engineering and Applied Sciences, 2017, pp. 3715-3718

[8] S. Kumar, M. Anamika Upadhyay, C. Gola, "Rainfall prediction based on 100 years of Meteorological data", IEEE, 2017, pp. 162-166.

[9] X. Xun, L. Mo, Y. Yu, "Discovery and Prediction of the Unused Land for Construction Based on Random Forest", Fifth International Conference on Agro-Geoinformatics, 2016.

[10] Y. C. Shiao, L. Liu, Q. Zhao, R. C. Chen, "Predicting Passenger Flow using Different Influence Factors for Taipei MRT System", IEEE 8th International Conference on Awareness Science and Technology (iCAST), 2017.

[11] S. Ye, X. Huang, Y. Teng, Y. Li, "K-Means Clustering Algorithm Based on Improved Cuckoo Search Algorithm and Its Application", IEEE 8th International Conference on Awareness Science and Technology, 2018, pp. 447-451.

[12] Z. Ya-Ling, W. Ya-nan, Y. Lil, "An Improved Sampling K-means Clustering Algorithm Based on MapReduce", IEEE 3rd International Conference on Big Data Analysis,2017.

[13] https://en.wikipedia.org/wiki/Machine_learning , Last Accessed on 05th Sept. 2018.

[14] https://en.wikipedia.org/wiki/Artificial_intelligence , Last Accessed on 05th.Sept. 2018.

[15] https://en.wikipedia.org/wiki/Linear_regression , Last Accessed on 04th Sept. 2018.

[16] https://en.wikipedia.org/wiki/Random_forest , Last Accessed on 05th Sept. 2018.

[17] https://en.wikipedia.org/wiki/K-means_clustering , Last Accessed on 05th Sept. 2018.

[18] B. R. Ranoliya, N. Raghuwanshi, S. Singh, "Chatbot for University Related FAQs", International Conference on Advances in Computing, Communications and Informatics (ICACCI), 2017.

[19] R. Ravi, "Intelligent Chatbot for Easy Web-Analytics Insights", International Conference on Advances in Computing, Communications and Informatics (ICACCI), 2017.