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. 7, Issue 3, March 2018

Proposing Regression Model for Predicting More Accurate Integer Sequences

Kartik Bhatia[1], Shourya Rastogi[2], Vibhor Agrawal[3], Mukesh Rana[4], Rachna Jain[5], Preeti Nagrath[6]

B.Tech, Dept. of CSE, Bharati Vidyapeeth College of Engineering, New Delhi, India[1]

B.Tech, Dept. of CSE, Bharati Vidyapeeth College of Engineering, New Delhi, India[2]

B.Tech, Dept. of CSE, Bharati Vidyapeeth College of Engineering, New Delhi, India[3]

B.Tech, Dept. of CSE, Bharati Vidyapeeth College of Engineering New Delhi, India[4]

Asst. Professor, Dept. of CSE, Bharati Vidyapeeth College of Engineering, New Delhi, India[5]

Asst. Professor, Dept. of CSE, Bharati Vidyapeeth College of Engineering, New Delhi, India[6]

ABSTRACT: Sequence Learning is an important branch of machine learning. Sequence learning helps us to solve mathematical problems by finding some pattern which previously was not possible earlier. Some sequences are answers to mathematical and physical problems in various area such as number theory. Knowing the next number may help to solve these problems and establish the validity of theories. To address this problem an effective approach is with machine learning using regression model. In this paper, we propose that how the regression plays a key role to generate more accurate sequences. This paper presents the idea to predict an integer sequence.

KEYWORDS: Sequence prediction, next item prediction, accuracy, regression, machine learning.

I. INTRODUCTION

Sequence prediction[1] is an important task with many applications. Let there be an integer of items (number) Z ={1, 2, ..., m}. A sequence s is an ordered list of items s = i1, i2, ...in, where i k $\in$ Z (1 $\leq$ k $\leq$ n). A set of training sequences is used to train the prediction model. Once trained, the model is used to perform sequence predictions. A prediction consists of predicting the next items of a sequence. This task has its applications in wide fields such as web page prefetching, consumer product recommendation, weather forecasting and stock prediction. Generally, the sequences are of two types which contains a linear relationship or non-linear relationship in between the numbers.A linear relationship is that which corresponds to a function that graphs the sequences in a straight line. All relationships other than the linear relationship fall under the category of non-linear relationship."A computer program is said to learn from experience E with respect to some class of tasks T and performance measure P if its performance at tasks in T, as measured by P, improves with experience E." It is used to develop a system that can automatically adapt and customize themselves to individual users e.g personalized news or mail filters.
It helps to discover new knowledge from databases (data mining) e.g. Market basket analysis. It has also the ability to mimic human and replace certain monotonous tasks which require some intelligence like recognizing handwritten characters. So it helps to develop systems that are too difficult/expensive to construct manually because they require specific detailed skills or knowledge tuned. In this paper, we present our approach with implementing regression[2] for linear and non-linear integer sequences. For, the linear model it would search a linear relationship between the input and output and so by plotting a graph we will be calculating the value from the graph to generate the output. For non-linear relations, we would first convert it in a linear relation and then proceed via the same previous method. Linear regression is a linear model in machine learning that assumes a linear relationship between the input variables and the single output variable. Specifically, the output can be calculated from a linear combination of the input. When there is a

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. 7, Issue 3, March 2018

single input, the method is referred to as simple linear regression and for multiple inputs, it is known as multiple linear regression.
There are basically four types of regression model used in machine learning.These are Simple Linear Regression model, Ordinary Least Squares, Gradient Descent, Regularisation.
The rest of the paper is organized as follows. Section 2 introduces related work done already in this field of sequence prediction. Section 3 is about proposed work and our approach towards the problem. Section 4 implies its applications and the future works.

II. RELATED WORK

Earlier the conventional approach was to add or subtract 1 or 2 from all the terms, and try looking it up again, Multiply all the terms by 2, or divide by any common factor, and try looking it up again or Look for a recurrence. But these are basically hit and trial and thus can be implemented by brute forcing only which is very inefficient. The other way which was used to predict sequences was used by Bruno Salvy and Paul Zimmermann written in 1994 was to manipulating of sequences to get a generating function in one variable. Gfun[3] is a classic Maple package for the manipulation of sequences to get generating functions by Bruno Salvy and Paul Zimmermann written in 1994 for getting of generating functions[4]. The gfun package consists of various functions for manipulating sequences linear recurrences or differential equations and various types of generating functions.
This is a powerful package which implements the use of generating functions for prediction of a sequence. But guessing a generating function is itself a very complex task as it involves complex mathematical operations on the elements thus it requires huge computational power. It also doesn't guarantee the desired accuracy for the consumed computingpower.
One of the most popular is Prediction by Partial Matching (PPM)[5]. It is based on the Markov property[6] and has inspired a multitude of other models such as Dependency Graph (DG)[7], All-K-Order-Markov (AKOM)[8], Probabilistic Suffix Tree (PST)[9] and Context Tree Weighting (CTW)[10]. Although, much work has been done to reduce the temporal and spatial complexity of these models. Besides, several compression algorithms have been adapted for sequence predictions such as LZ78 and Active Lezi. Moreover, machine learning algorithms such as neural networks and sequential rule mining have been applied to perform sequence prediction.
However, these models suffer from some important limitations. First, of all, it assumes the Markovian hypothesis that each event individually depends on the previous events. If this hypothesis does not hold, prediction accuracy using these models can severely decrease. Second, all these models are built using only part of the information contained in training sequences. Thus, these models do not use all the information given in training sequences to perform predictions and reduces its accuracy. For instance, Markov models typically consider only the last k items of training sequences to perform a prediction, where k is the order of the model. This often induces a very high state complexity, thus making them impractical for many real-life applications.
To tackle this problem another approach is to fetch sequential rules from a trained sequence of data then use these rules to make predictions for new sequences by proposing to use partially-ordered sequential rules instead of standard sequential rules.
However, a problem with this idea of using patterns for predictions is that you need to set the minimum support. If you set this parameter to 25% for example, then you will miss all the rare cases that have a support below 25 %, and thus the prediction for these rare cases may not be good. But rare cases are also important if you want your model to have a high prediction accuracy.
Another approach was to use a model named Compact Prediction Tree (CPT)[11]. CPT works with the proposal of a prediction model which mainly compresses training sequences without information loss by exploiting similarities between subsequences. It depends on a tree type structure and a has a more complex prediction algorithm which offers more accurate predictions than any other prediction models.
However, a drawback of CPT is that of spatial complexity and requires a higher prediction time.

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. 7, Issue 3, March 2018

III. PROPOSED WORK

I.1 Data

Our implementation will be working only on the linear type of data so it is taken from the OEIS[12] and first of all we will be separating out the linear and non-linear data.If found a non-linear data we will be converting it into a linear type of data and then process.

I.2 Applying Regression model

For each given set of data values, a curve is plotted out and the next value is predicted along the curve.Once data is separated out from the training set we will be applying one of the regression models on it and selecting out the best model and further making a change so to increase its accuracy and efficiency.
Lastly training that model with the data will generate a more accurate, precise and efficient model.
Let y is function of x such that y=f(x).Where f(x) is any linear function which has a polynomial degree of 1.
So, y can be defined as y=a1*x+a2*x+a3*x.....an*x+c.

Where 'c' is any constant and a1,a2,a3...are the coefficients of x.Whose graph can be plotted as

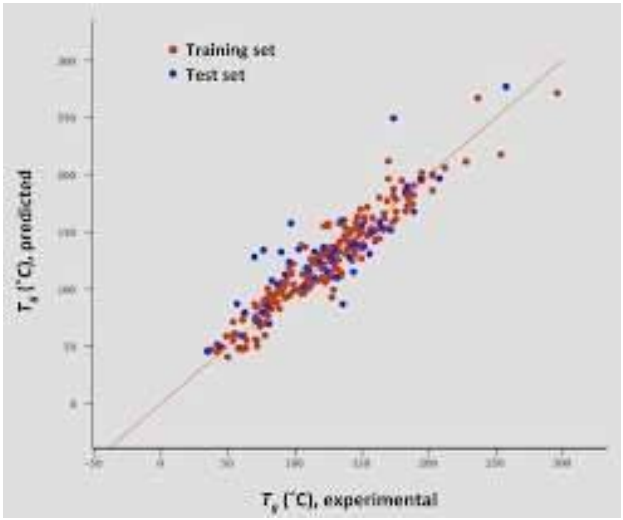

Fig 1: Curve plotted for a set of data point P versus the experimental analysis

In the above figure we have taken experimental values in the x-axis and predicted values in the y- axis. We have plotted a set of data points of the sequences which is given for training and testing the machine, here orange coloured points are the training set points and blue coloured points are the test set points. After plotting, a straight line is made which is passing through this points and further we predict that the next integer sequence number will lie on that line or nearby it.

Fig 2:Demo display of working model

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. 7, Issue 3, March 2018

The above figureexplains that the linear data which is sorted, is passed in the machine and after finding the corresponding values and compared with the trained data set, a predicted sequence of numbers is derived which is the output .

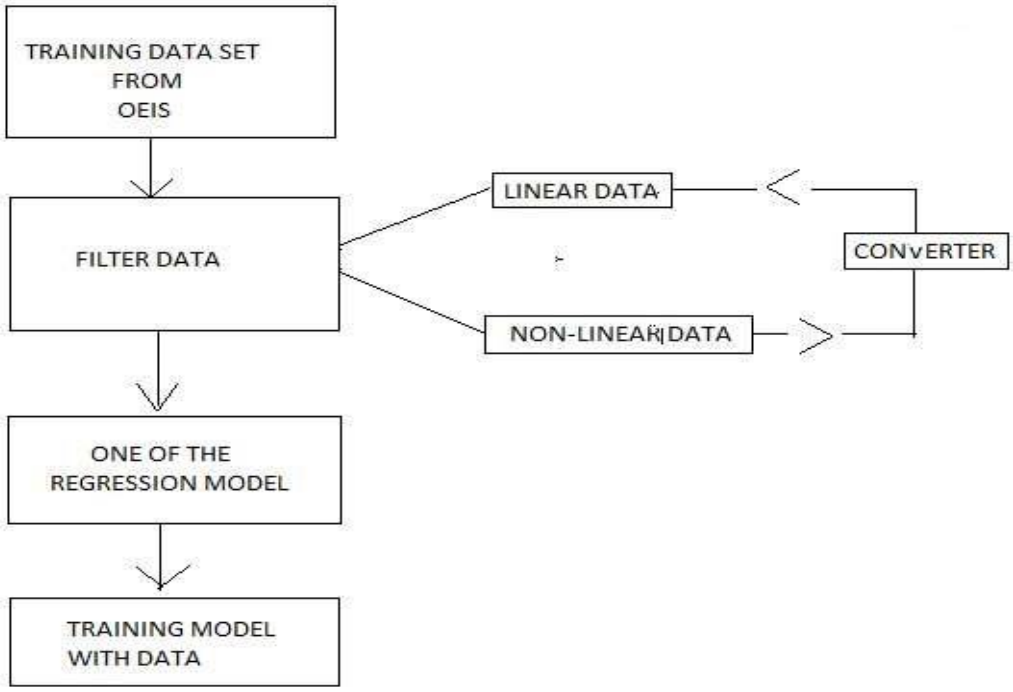

Fig 3:The flowchart depicting sequence prediction for a given sequence

The above figure illustratestraining a sequence prediction model using some previously seen sequences and touse a trained sequence prediction model to perform prediction for new sequencesin this we have taken data set for OEIS and after that data is filtered to linear data. Linear data is passed into the simple regression model for training.

II. APPLICATIONS AND CONCLUSION

The scope of this project is not only limited to the field of integers but could be extended to wide fields such as to predict next character of the string sequence.
This can also be used to solve complex mathematical and physical problems. Some sequences are answers to mathematical and physical problems in various area such as number theory. Knowing the next number may help to solve these problems and establish the validity of theories. We are sometimes unable to calculate these by conventional means, either because we are not sure about how to model the problem, or because it goes beyond our technical abilities.

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. 7, Issue 3, March 2018

We can also use sequence prediction in cryptography which can help us avoid the use of prime-ness as a cryptography[13] tool which is a very complex task and requires brute forcing. Further to add this can also be used in stock prediction and weather forecasting.

REFERENCES

[1] Philippe Fournier-Viger, "An Introduction to Sequence Prediction," link-http://data-mining.philippe-fournier-viger.com/an-introduction-to-sequence-prediction/, 2016-06-23.
[2] Abhijit Dasgupta, Yan V. Sun, Inke R. Konig, Joan E. Bailey-Wilson, and James D. Malley, "Regression-Based and Machine Learning Methods," vol. 35, pp. 7-11, 2012.
[3] Bruno Salvy, Paul Zimmermann: GFUN, "a maple package for the manipulation of generating and holonomic functions in one variable," no. 14, 1992.
[4] Donald E. Knuth, "The Art of Computer Programming,"vol. 1, pp. 87-95, 1997.
[5] Cleary, J., Witten, I., "Data compression using adaptive coding and partial string matching," vol. 24, no. 4, pp. 413–421, 1984.
[6] Markov, A. A., "Theory of Algorithms," vol. 42, no.4, 1954.
[7] Thomas Zimmermann, Nachiappan Nagappan, "Predicting defects using network analysis on dependency graphs in Software Engineering," ICSE '08, 2008.
[8] Begleiter, R., El-Yaniv, R., Yona., G.,"On prediction using variable order Markov models," vol. 22, pp. 385-421, 2004.
[9] Alexis Gabadinho, Gilbert Ritschard,"Analyzing State Sequences with Probabilistic Suffix Trees," vol. 72, no. 3, pp. 1-39, 2016.
[10] Willems; Shtarkov; Tjalkens (1995), The Context-Tree Weighting Method: Basic Properties, 41, IEEE Transactions on Information Theory
[11] Ted Gueniche, Philippe Fournier-Viger1, Rajeev Raman and Vincent S. Tseng: CPT+: Decreasing the time/space complexity of the Compact Prediction Tree published in ACM SAC in the year 2011.
[12] N.J.A Solane, "OEIS- On-Line Encyclopedia of Integer Sequences," 1964.
[13] Ronald L. Rivest,"Cryptography and machine learning in Theory and Application of Cryptology," pp. 427-439, 1994.