

Imputation and Regression Techniques to Estimate Missing Values Using Big Data Analytics

Prof. Deepali Patil

Assistant Prof, H.O.D. Department of Information Technology
University of Mumbai
Shree L.R. Tiwari College of Engineering
Mira Road, Maharashtra
deepali.patil@slrtce.in

Abstract - Today, many organizations are collecting, storing, and analyzing massive amounts of data. This data is commonly referred to as “big data” because of its volume, the velocity with which it arrives, and the variety of forms it takes. A key to deriving value from big data is the use of analytics. Collecting and storing big data creates little value; it is only data infrastructure at this point. It must be analyzed and the results used by decision makers and organizational processes in order to generate value. Situations may occur where sometimes data is missing. But it becomes problematic if the data is related to medical Sector. And processing of such data becomes difficult.

Hence, we have proposed algorithm to find and predict Missing Value from historical values. This algorithm will be used only for numeric values because we have used linear regression and polynomial regression Technique to predict the value. We extend this prediction by creating new K-NN classifier based on map reduce functionality to impute missing value.

I INTRODUCTION

For large quantity of information, it's virtually not possible for human analysts to derive meaningful conclusions in a very short timeframe. Thus data processing techniques are looked upon as tools that may be used to automate the method of knowledge discovery and outline relationships and patterns of resemblance given a totally random and raw data set. The bulk of information collected for analysis is unsupervised. This provides rise to the necessity of effective techniques that can process such unsupervised data sets and convert what might seem to be fully random and meaningless into something sensible and valuable.

We explore the rule to seek out missing value for numeric data set. It focuses on regression model missing value estimation technique, Sridevi.S et. Al [3] do autoregressive-model-based

missing value estimation technique but they focus on time series data. But they have some disadvantage - like it is only for time series data and in huge data set time series data is extremely less. Time series data set means that in historical data we've one field like time so data is accessible from all time series.Example: Temperature, weather etc.

Thus we propose algorithm for numeric data. Here we've used regression toward the mean and polynomial regression model to predict future value or missing value. For find missing value we apply regression toward the mean on input. We've also proposed a MapReduce approach. Hadoop[2][8] is an open source application that runs on a distributed computing environment and supports processing of high volume data intensive applications. it's evolved from The Google file system. Hadoop uses MapReduce[8]programming style that

provides it the flexibleness and capabilities required to process petabytes of information. Since Hadoop uses the MapReduce paradigm it achieves its goal of process by downscaling the given data set and consequent integration of the data processed one by one at separate nodes that are networked to form a cluster. Hence the computational demands are met collectively by all the nodes that form the cluster and this gives rise to extend in efficiency that is in harmony with the low costs involved in setting up the cluster.

Map Reduce technique enables the k-Nearest neighbor technique to deal with large-scale problems. Without such a parallelization, the application of the k-NN algorithm would be limited to small or medium data, especially when low runtimes are a need. The proposed scheme is an exact parallelization of the k-NN model, so that, the precision remains the same and the efficiency has been largely improved.

II LITERATURE REVIEW

S.Sridevi, Dr. S.Rajaram, C.Parthiban, S.SibiArasan and C.Swadhikar [3] designed autoregressive-model-based missing value estimation method (ARLSimpute). This algorithm is used to find missing value for time series data set. So at any point of time some data is missing or if the whole time point is missing so we can find using this algorithm. They had used linear prediction technique and quadratic prediction technique to find missing values. Also they predict future data from historical data.

Thirumagal R, Deepali A Patil [4] builds two algorithms KNN impute and ARL impute. These are used to impute and find missing values from data set. Here KNN gives best result if value of k e is between 10 to 20. This KNN impute algorithm is used to find only one missing data from only one column

but ARL imputation algorithm is used for finding many missing values from one column and from entire time point.

Qianya Zhang, Ashfaqur Rahman, and Claire D'Este [5] evaluated imputation technique performance in sensor field as sensor faults and sensor communication error some time some readings are missing . They considered two methods for this (a) the missing value only in prediction phase (b) Missing value in introduction phase as well as Prediction phase. They focused on two words "Impute" and "Ignore" for missing value. Impute means missing value estimated using imputation to existing value in data set through Prediction. Ignore means avoid the missing value if it is not useful.

Y.Tao, D.Papadias and X.Lian builds the Local Least Square Imputation algorithm (LLSimpute). They used KNN model Process that selected neighboring value for missing data and predicate from test data.

III PROPOSED METHOD

The proposed algorithm is based on numeric data set and here we are taking the data : " Synthetic control chart" Example 1-20 Normal

21-40 Cyclic

41-60 Increasing trend

61-80 Decreasing trend

81-100 Upward shift

101-120 Downward shift. According to this we have taken many samples. We divide data set in 2 parts (1) training data (2) testing data. Training data is the data which is not needed in prediction, means those attributes are not needed in prediction , remove it and take it in this data set, and almost all data set have 80% training data. Testing data is the data which is required for prediction analysis. So we only find missing value in this data. In this very few attribute is available i.e. only 20% data is come under training data. This technique used to predicate future value from historical data.

Proposed algorithm is given below.

Proposed Algorithm

1. Start
2. Load Data Set
3. Divide the data in to Training Data and Testing Data.
4. Load the Training Data.
5. Convert testing data in 2 D matrix Format.
6. Estimate the Missing Value (MV) from Testing Data.
7. Select that Row (R).
8. Get all column value for R.
9. LV=Last value
10. LVC=Last value column number
11. For (n = starting column (1); n < MN Column; n++)

$$N_{sum} = \sum_{i=1}^n (\chi_n - \chi_{n-1}) \quad (1)$$

12. Get Previous value of MV and stored in PV.

$$N_{missing} = N_p \pm \frac{N_{sum}}{n} \quad (2)$$

Where n = last value place?

14. Combine Training Data and Testing Data.

15. End Session

The k-Nearest Neighbor classifier is one of the most simplest algorithm in data mining because of its effectiveness and simplicity. But it lacks the scalability to manage big datasets. The main issue for dealing with large-scale data are runtime and memory consumption.

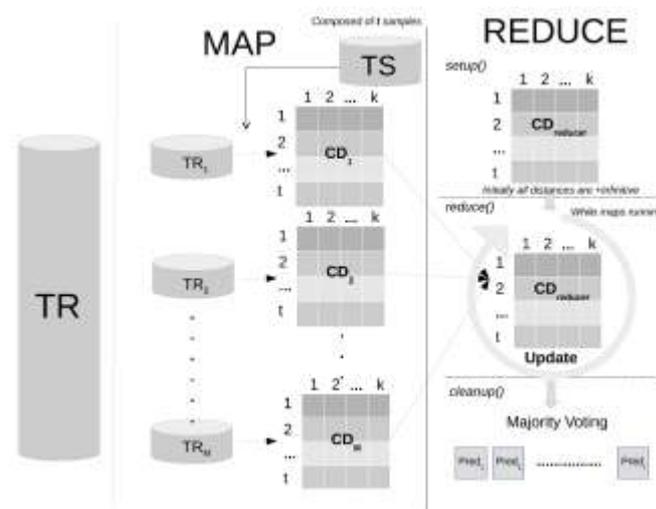


Fig1: Map reduce technique.

Algorithm for Map function

K-NN Mapper

```
{ Create list to maintain data points in the testing dataset
Test List= new TestList
Load file containing testing dataset
Load testfile
Update list with data points from file
TestList<=testfile
Open file contacting training dataset
Open trianfile
Load training data points one at a time and compute distance
with every testing data point
Distance(traindata,testdata)
Write the distance of test data points from all the training data
points with their respective class labels in ascending order of
disatances
Testfile<=test data(dist, label)
```

```
Call reducer  
}
```

Algorithm for Reduce function

```
K-NN Reducer  
{  
Load the value of k  
Load testfile  
Open testfile  
Load test datapoints one at a time  
Read testdatapoint  
Initialize counters for all calss kabela  
Set counter=0  
Look through top k distance for the respective test data point  
and increment the corresponding class label counter.  
For i=0 to k  
    Counter++  
    Assign the class label with the highest count for the  
testdatapoint  
    Test datapoint = class label(counter max value)  
    Update output file with classified test data point  
    outFile= outFile +testdatapoint  
}
```

- [7] Choong, Miew Keen, Maurice Charbit, and Hong Yan. "Autoregressive-model-based missing value estimation for DNA microarray time series data." *Information Technology in Biomedicine, IEEETransactions on* 13.1 (2009): 131-137.
- [8] A. Fernandez, S. R ´ ´io, V. Lopez, A. Bawakid, M. del Jesus, J. Ben ´ ´itez, and F. Herrera, "Big data with cloud computing: An insight on the computing environment, mapreduce and programming frameworks," *WIREs Data Mining and Knowledge Discovery*, vol. 4, no. 5, pp. 380–409, 2014.

IV CONCLUSION

In recent years due to large data sets and handling such large database is becoming critical. Hence research in big data is increasing day by day. And also Organizations have large Database like Health Organization, Social Media, Shopping site like amazon, flipkart. And all these data comes from users. So there is possibility to have some missing value. In existing system imputation was done on time series data and small scale data. Our aim is to predict such missing values from big data and performs analytics on that so that in every sector we will get accurate results.

REFERENCES

- [1] J. Dean and S. Ghemawat, "Mapreduce: simplified data processing on large clusters," *Communications of the ACM*, vol. 51, no. 1, pp. 107– 113, Jan. 2008.
- [2] <http://www-01.ibm.com/software/data/bigdata/what-is-big-data.html>
- [3] Sridevi, S., et al. "Imputation for the analysis of missing values and prediction of time series data." *Recent Trends in Information Technology (ICRTIT)*, 2011 International Conference on. IEEE, 2011.
- [4] Thirumahal, R., and Deepali A. Patil. "KNN and ARL Based Imputation to Estimate Missing Values." *Indonesian Journal of Electrical Engineering and Informatics (IJEI)* 2.3 (2014): 119-124.
- [5] Zhang, Qianyu, Ashfaqur Rahman, and Claire D'Este. "Impute vs. ignore: Missing values for prediction." *Neural Networks (IJCNN)*, The 2013 International Joint Conference on. IEEE, 2013.
- [6] Tao, Yufei, Dimitris Papadias, and Xiang Lian. "Reverse kNN search in arbitrary dimensionality." *Proceedings of the Thirtieth international conference on Very large data bases- Volume 30. VLDB Endowment*, 2004.