INTEGRATION OF BIG DATA WITH ML- A NOVEL APPROACH TO FOREKNOW ELECTRICITY GENERATION

Aditi Gupta (Assistant Professor)
Department of computer science, D. A. V. College for Boys, Amritsar (143001), Punjab, India
Main author: Aditi Gupta , City: Amritsar 143001,Punjab

Abstract

Big data has increased the demand of information management specialists so much so that Software AG, Oracle Corporation, IBM, Microsoft, SAP, EMC, HP and Dell have spent more than $15 billion on software firms specializing in data management and analytics. In 2010, this industry was worth more than $100 billion and was growing at almost 10 percent a year: about twice as fast as the software business as a whole.

1. Introduction

This study uses a very different amalgamation of Big Data and Machine Learning to predict power generation in United States. Big Data is employed to process large amount of unstructured and semi-structured datasets complemented with noise .This processed data is then fed to ML model to predict the electricity generation which then forecasts the results based on the past datasets. This peculiar combination gives very efficient and accurate results.

2. Literature Survey

1. "The model can forecast future power generation based on the collected data, and our test results show that the proposed system can predict the required power generation close to 99% of the actual usage." This statement has been claimed after the experiment conducted on the datasets of years 1980-2014.
2. "MAPE percentage was calculated to be 4.13% for total generation forecast and the individual forecast values are in the range of 4–9% for all the states." The claim is evident from the calculations made for total generation of U.S and actual generation of few states.

3. Methodology

In this work, efficient electricity forecasting is built using the ML approach with Big Data to overcome the challenges related to large datasets. The proposed framework is designed not only to build an effective forecasting system, but also to solve the problems lems have noise, using distributed algorithms in the form of MapReduce.

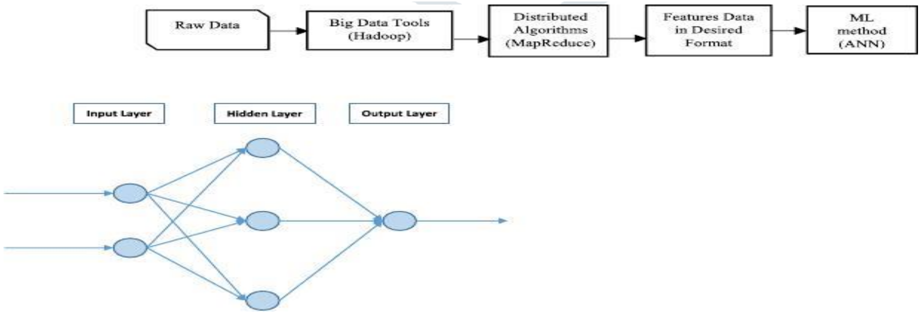

Fig 1: Artificial neural network processing

Initially,the raw data is stored in HDFS inside the Hadoop cluster. HDFS stores files in a distributed fashion, and it also replicates data blocks in different nodes (for this work the replication factor was set to default value of 3). Hadoop breaks the data into chunks or blocks to be stored inside HDFS. The data can be divided into blocks of 64, 128, and 256 MB. In this work, default block size of 64 MB is chosen. The data are first divided into blocks and then placed into HDFS; later the replication is performed. The reason for storing data in a distributed format is to perform parallel processing and computation of large data, while increasing reliability, flexibility, and scalability. Then we applied MapReduce, a low level language to retrieve desired features from data. We have implemented Mapper and Reducer algorithms in MapReduce to perform their tasks. The Mapper function tells the cluster which data points are required to be retrieved, and then the Reducer acquires and aggregates all the data, and converts it to a suitable format . The Hadoop cluster contains one master and several slave nodes (NameNode acts as master and data nodes act as slaves.

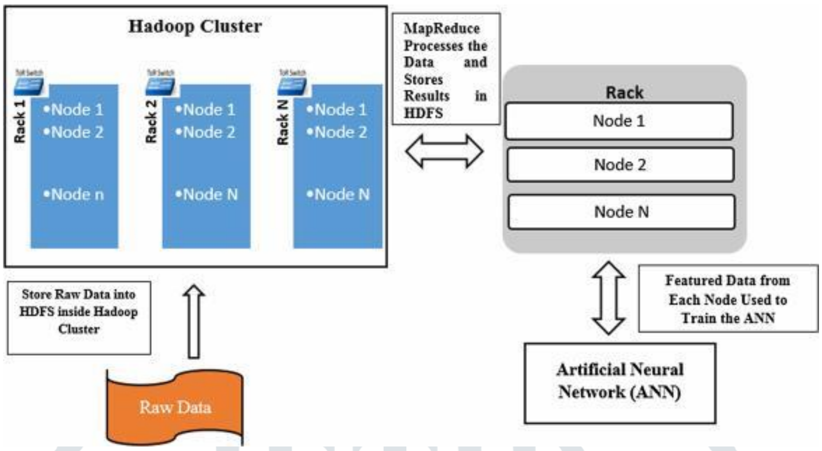
Hadoop Cluster
Rack 1
•Node 1
•Node 2
•Node n
Rack 2
•Node 1
•Node 2
•Node N
Rack N
•Node 1
•Node 2
•Node N
MapReduce Processes the Data and Stores Results in HDFS
Rack
Node 1
Node 2
Node N
Store Raw Data into HDFS inside Hadoop Cluster
Raw Data
Featured Data from Each Node Used to Train the ANN
Artificial Neural Network (ANN)

Fig 2: Big data cluster with Machine learning

The output data in structured format stored in HDFS is retrieved for training the BPNN, which is an important part of the framework. Data is divided into two sets: 90% used for training the network and the remaining 10% for testing the network. For each prediction; In the input layer, there are 12 nodes; in the hidden layer, 6 nodes; and in the output layer one node (12-6-1). The size of the input layer contains the number of features in the data. Before setting the number of input nodes to 12, the forecasting results are evaluated using the 3rd, 4th, 6th and 8th input nodes. After this evaluation, the generation of next month is forecasted by the past 12 months' generation data, which has been included into the network. Hence, the algorithm outputs optimal results for the past 12 months as input into 12 nodes. The algorithm can recognize the pattern very well if the entire year is used. The size of output layer is also determined in a similar manner. BPNN can be run in two different ways: ML mode and Regression mode. ML mode determines the output as class label, and the re-gression mode returns values (e.g. predicting price). In this work, BPNN runs on regression mode and the output layer has a single node. There is one hidden layer with 6 nodes.

4. Results and Conclusion

The output data in structured format stored in HDFS is retrieved for training the BPNN, which is an important part of the framework. Data is divided into two sets: 90% used for training the network and the remaining 10% for testing the network. For

each prediction; In the input layer, there are 12 nodes; in the hidden layer, 6 nodes; and in the output layer one node (12-6-1). The size of the input layer contains the number of features in the data. Before setting the number of input nodes to 12, the fore-casting results are evaluated using the 3rd, 4th, 6th and 8th input nodes. After this evaluation, the generation of next month is fore-casted by the past 12 months' generation data, which has been included into the network. Hence, the algorithm outputs optimal results for the past 12 months as input into 12 nodes. The algorithm can recognize the pattern very well if the entire year is used. The size of output layer is also determined in a similar manner. BPNN can be run in two different ways: ML mode and Regression mode. ML mode determines the output as class label, and the regression mode returns values (e.g. predicting price). In this work, BPNN runs on regression mode and the output layer has a single node. There is one hidden layer with 6 nodes.

5. References

1 Electricity datasets from Energy Information Administration, available on web link: http://www.eia.gov/beta/api/qb.cfm?category=0, Web – 20 Jan, 2014.

2 Wikipedia contributor (2014, Jan 10), Big data (version ID: 409751071) [online], available: http://en.wikipedia.org/wiki/Big_data.

3 Wikipedia contributor (2013, Nov 23), Machine Learning (version ID: 175534373) [online], available: http://en.wikipedia.org/wiki/Machine_learning, 2013, Web – 12 Jan, 2014.

4 T. Senjyu, H. Taakara, K. Uezato, T. Funabashi, One-hour-ahead load forecasting using neutral network, IEEE Trans. Power Syst. 17 (1) (2002) 113–118.

5 K. Orwig, M. Ahlstrom, V. Banunarayanan, J. Sharp, J. Wilczak, J. Freedman, S. Haupt, J. Cline, O. Bartholomy, H. Hamann, B. Hodge, C. Finley, D. Nakafuji, J. Peterson, D. Maggio, M. Marquis, Recent trends in variable generation forecasting and its value to the power system, IEEE Trans. Sustain. Energy (2015) 1–10.

6 C. Wan, Z. Xu, P. Pinson, Z. Dong, K. Wong, Probabilistic forecasting of wind power generation using extreme learning machine, IEEE Trans. Power Syst. 29 (3) (2014) 1033–1044.

7 Jeff Heaton, Introduction to Neural Network for Java, 2nd edition, Heaton Research, ISBN 1604390085, October 2008.

8 MapReduce tutorial (2015, Dec 18), The Apache Software Foundation [online], available: https://hadoop.apache.org/docs/r1.2.1/mapred_tutorial.html.

9 A. Setaiwan, I. Koprinska, V.G. Agelidis, Very short-term electricity load demand forecasting using support vector regression, in: International Joint Conference on Neural Network, Atlanta, Georgia, USA, vol. 1, 2009, pp. 2888–2894.

10 H. Wang, S. Zhu, J. Zhao, G. Li, An improved combined model for the electricity demand forecasting, in: International Conference on Computational and Information Sciences, vol. 6, 2010, pp. 107–111.

11 J.N. Fidalgo, M.A. Matos, Forecasting Portugal global load with Artificial Neural Network, LNCS 4669 (2) (2007) 728–737.

12 P.F. Pai, W.C. Hong, Forecasting regional electricity load based on recurrent support vector machines with genetic algorithms, Electr. Power Syst. Res. 74 (3) (2005) 417–425.

13 P.C. Chang, Y.W. Wang, C.H. Liu, Fuzzy Delhi and Backpropagation model for sales forecasting in PCB industry, Expert Syst. Appl. 30 (4) (2006) 715–726.

14 M.C. Su, C.W. Liu, S.S. Stay, Neural Network based Fuzzy Model and its application transient stability prediction in power system, IEEE Trans. Syst. Man Cybern., Part C, Appl. Rev. 29 (1) (1999) 149–157.

15 R.H. Liang, Application of gray relation analysis to hydroelectric generation scheduling, Int. J. Electr. Power Energy Syst. 21 (5) (1999) 357–364. [25] X. Wu, X. Zhu, G. Wu, W. Ding, Data mining with Big Data, IEEE Trans. Knowl. Data Eng. 26 (1) (Jan 2014).