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. 8, Issue 5, May 2019

Landslide and Flood Prediction Using Machine Learning Spark Framework

Mihir Sandeep Kungulwar, Harsh Sanjeev Mishra, Shahzer Ahmed Khattak, Uzair Nazir Bhat

Department of Computer Engineering, Sinhgad College of Engineering, Pune, India

ABSTRACT: Landslide and flood prediction is a major problem in recent years. The statistical analysis of each region is essential to estimate the input value related to the project Analysis of engineering structures and also for crop planning. To understand normality, daily rainfall data over a period of 30 years are used. In addition, different route position formulas are used to evaluate the return period of monthly, seasonal and annual rains. This analysis will provide useful information for the rainfall and flood prediction. We uses the random forest algorithm to classify the regional rainfall patterns.in order to obtain the meteorological information on the internet. The meteorological data such as rainfall, temperature and humidity at the weather station will be crawled from the web. After pre-processing metrological data this paper applies random forest on the platform of apache spark to classify results.

KEYWORDS: Big Data, Machine Learning, Random Forest, Spark

I. INTRODUCTION

Big data analytics is a technology which helps turn hidden insights in big data into business value by using advanced analytics techniques in order to support better business decisions, as the hottest new practice in Business Analytics today. According to some surveys, about 80% of CEOs or executive teams view that big data analytics initiatives have the potential to drive business value such as creating new revenue streams, improving operational efficiency or cutting cost, and over 80% of the participant organizations do ongoing projects. However, according to another survey, 55% of big data projects do not get completed, and many others fall short of their business objectives.
Additionally, the quality of big data analytics can only be as good, or as bad, as the quality of the big data it uses. The quality attributes of big data, together with relationships between data, should therefore be defined in a big data model. With a good quality big data model, big data analytics can accurately identify important business concerns, trend opportunities, and useful business insights, which, in turn, can lead to prediction system. Yet, no guidelines are available for how to develop a high-quality big data model in a systematic and rational way.

II. RELATED WORK

Literature survey is the most important step in any kind of research. Before start developing we need to study the previous papers of our domain which we are working and on the basis of study we can predict or generate the drawback and start working with the reference of previous papers.

In this section, we briefly review the related work on Prediction system and their different techniques.

Informatics and Capgemini, The Big Data Payoff: Turning Big Data into Business Value, 2016: The same amount was created in every two days in 2011, and in every ten minutes in 2013.Assess the business value and benefits that enterprises are seeking and realizing from Big Data.

S. Supakkul, L. Zhao and L. Chung, "GOMA:Supporting Big Data Analytics with a Goal-Oriented Approach," IEEE BigData Con.,2016. pp. 149-156: The real value of Big Data lies in its hidden insights, but the current focus of the Big Data community is on the technologies for mining insights from massive data, rather than the data itself. The biggest challenge facing industries is not how to identify the right data, but instead, it is how to use

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. 8, Issue 5, May 2019

insights obtained from Big Data to improve the business. To address this challenge, we propose GOMA, a goal-oriented modeling approach to Big Data analytics. Powered by Big Data insights, GOMA uses a goal-oriented approach to capture business goals, reason about business situations, and guide decision-making processes. GOMA provides a systematic approach for integrating two types of the resulting insight from data analytics to goal-oriented reasoning and decision-making processes: descriptive insights are the ones that describe the current state (e.g., the current customer retention rate) and predictive insights are the ones that predict likely future phenomena by inference from the data (e.g., customers who are likely to defect). To aid in the description and illustration of the GOMA approach, a retail banking churning scenario is used as a running example throughout this paper.

A. Davidson and A. Or. "Optimizing shuffle performance in spark." University of California, Berkeley-Department of Electrical Engineering and Computer Sciences, Tech. Rep, 2013:
Spark has been established as an attractive platform for big data analysis, since it manages to hide most of the complexities related to parallelism, fault tolerance and cluster setting from developers. However, this comes at the expense of having over 150 configurable parameters, the impact of which cannot be exhaustively examined due to the exponential amount of their combinations. The default values allow developers to quickly deploy their applications but leave the question as to whether performance can be improved open. In this work, we investigate the impact of the most important tunable Spark parameters with regards to shuffling, compression and serialization on the application performance through extensive experimentation using the Spark-enabled Marenostrum III (MN3) computing infrastructure of the Barcelona Supercomputing Center. The overarching aim is to guide developers on how to proceed to changes to the default values. We build upon our previous work, where we mapped our experience to a trial-and-error iterative improvement methodology for tuning parameters in arbitrary applications based on evidence from a very small number of experimental runs. The main contribution of this work is that we propose an alternative systematic methodology for parameter tuning, which can be easily applied onto any computing infrastructure and is shown to yield comparable if not better results than the initial one when applied to MN3; observed speedups in our validating test case studies start from 20%. In addition, the new methodology can rely on runs using samples instead of runs on the complete datasets, which render it significantly more practical.

TONY HALL*,HAROLD E. BROOKS AND CHARLES A. DOSWELL" Precipitation Forecasting Using a Neural Network": The neural network, using the Eta model input and the upper air probes, was developed for probability of precipitation (PoP) and quantitative precipitation forecast (QPF) for Dallas-Fort Worth, Texas, area. The biennial forecasts were verified with a network of 36 rain gauges. The resulting predictions were extraordinarily strong, with over 70% of PoP forecasts of less than 5% or more than 95%.

Pengcheng Zhang1, Lei Zhang, Hareton Leung, Jimin Wang" A deep-learning based precipitation forecasting approach using multiple environmental factors": The proposed Approach not only you can learn hierarchical representation of raw data using a highly generalized form, completely undermining the information hidden in the original data, but even more Accurate description of the rule behind the different types of Large time series of data. Seven types of environmental factors, which are much related to precipitation, are used as input the vector and subsequent 24-hour precipitation are used as output Vector. A series of dedicated experiments with data from the Zunyi area from the province of Guizhou is carried out to validate the profitability andUtility of the model. We also compare deep-based learning focus on other traditional approaches to machine learning.Experimental results show that the proposed approach canimprove the accuracy of rainfall forecasts.

III. PROPOSED APPROACHES

Lot of work has been done in this field because of its extensive usage and applications. In this section, some of the approaches which have been implemented toachieve the same purpose are mentioned. These works are majorly differentiatedby the algorithm for prediction systems.
As my point of view when I studied the papers the issues are related to predictionsystems. The challenge is to addressing flood prediction problem is based on theenvironment factors dataset.

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. 8, Issue 5, May 2019

A. *System Architecture:*

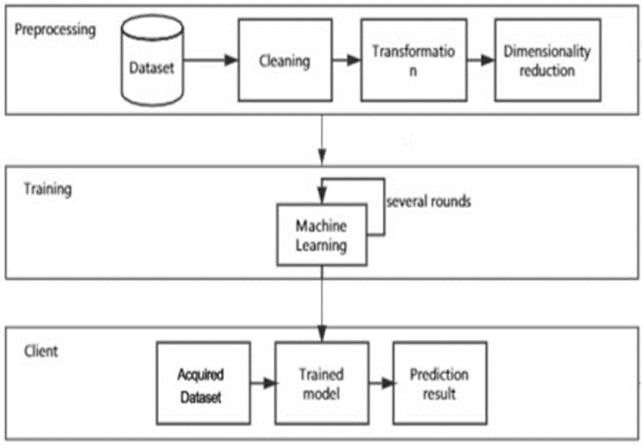

Fig. System Architecture

B. *Algorithm*

Random forests is a machine learning algorithm. The basic premise of the algorithm is that building a small decision-tree with few features is a computationally process.
The algorithm works as follows: for each tree in the forest, we select a bootstrap sample from S where S(i) denotes the ith bootstrap. We then learn a decision-tree using a modified decision-tree learning algorithm.
The algorithm is modified as follows: at each node of the tree, instead of examining all possible feature-splits, we randomly select some subset of the features f subset F:
Where F is the set of features. The node then splits on the best feature in f rather than F. In practice f is much, much smaller than F. deciding on which feature to split is oftentimes the most computationally expensive aspect of decision tree Learning. By narrowing the set of features, we drastically speed up the learning of the tree.

IV. RESULTS AND DISCUSSION

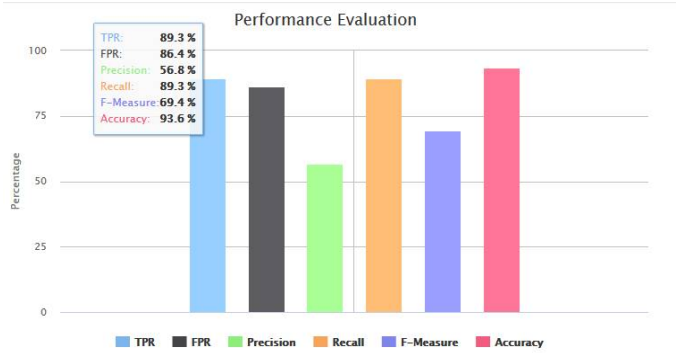

Fig. Analysis Graph

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. 8, Issue 5, May 2019

Parameters	Percentage
TPR	85.1
FPR	78.7
Precision	60.6
Recall	85.1
F-Measure	78.8
Accuracy	93.4

V. CONCLUSION

In this Paper, Natural things are outside human control, but destruction can be reduced if the prediction mechanisms are Done in advance Researchers from all over the world are having A big step to develop early prediction mechanisms for these natural risks. It was difficult to use traditional mathematics Analysis methods. So random forest helped us get results using historical data. In this way, the study concludes that random forest has proven to be an efficient technique to predict landslides during forecasting early rain.

REFERENCES

[1] Y. Y, Chen, A. J, Cheng, and W. H. Hsu, "Travel recommendation by mining people attributes and travel group types from community-contributed photos," IEEE Transactions on Multimedia, vol. 15, no. 6, pp. 1283-1295, Oct. 2015.
[2] P. Kefalas, P. Symeonidis, and Y. Manolopoulos, "A graph-based taxonomy of recommendation algorithms and systems in LBSNs," IEEE Transactions on Knowledge and Data Engineering, vol. 28, no 3, pp.604-622, Mar. 2016.
[3] P. Peng, L. Shou, K. Chen, G. Chen, and S. Wu, "KISS: knowing camera prototype system for recognizing and annotating places-of-interest,"IEEE Transactions on Knowledge and Data Engineering, vol. 28, no 4, pp.994-1006, Apr. 2016.
[4] Personalized Travel Sequence Recommendation on Multi-Source Big Social Media Shuhui Jiang, XuemingQian *, Member, IEEE, Tao Mei, Senior Member, IEEE and Yun Fu, Senior Member, IEEE
[5] X. Wang, Y. L. Zhao, L. Nie, Y. Gao, W. Nie, Z. J. Zha, and T. S. Chua, "Semantic-based location recommendation with multimodal venue semantics," IEEE Transactions on Multimedia, vol. 17, no. 3, pp. 409-419, Mar. 2015.
[6] S. Jiang, X. Qian, J. Shen, Y. Fu, and T. Mei, "Author topic model based collaborative filtering for personalized poi recommendation," IEEE Transactions on Multimedia, vol. 17, no. 6, pp. 907–918, 2015.
[7] Q. Hao, R. Cai, X. Wang, J. Yang, Y. Pang, and L. Zhang, "Generating location overviews with images and tags by mining usergenerated travelogues," in Proceedings of the 17th ACM international conference on Multimedia. ACM, 2009, pp. 801–804.
[8] Q. Yuan, G. Cong, Z. Ma, A. Sun, and N. M. Thalmann, "Time-aware point-of-interest recommendation," in Proc. SIGIR, 2013, pp. 363-372.
[9] J. D. Zhang and C. Y. Chow, "Spatiotemporal sequential influence modeling for location recommendations: a gravity-based approach," ACM Transactions on Intelligent Systems and Technology, vol. 7, no. 1, pp. 11, Jan. 2015.
[10] J. D. Zhang and C. Y. Chow, "Point-of-interest recommendations in location-based social networks," in Proc. SIGSPATIAL, 2016, pp. 26-33.