ISSN: 0974-5823 Vol. 6 No. 3 December, 2021

International Journal of Mechanical Engineering

PREDICTION OF RAINFALL ACROSS VARIOUS REGIONS USING DEEP LEARNING BASED LSTM

S. Karuppusamy*
[1]Associate Professor, Department of CSE, Nandha Engineering College, Erode, Tamil Nadu, India.
Mail Id: sksamymsc@gmail.com

R.Arshath Raja
Senior Associate, Research and Publications, ICT Academy, IIT Madras Research Park, Chennai, Tamil Nadu 600113, India.

Yuvaraj
Manager, Research and Publications, ICT Academy, IIT Madras Research Park, Chennai, Tamil Nadu 600113, India.

C.Selvin Thanuja
Associate Professor, Department of Pharmaceutical Chemistry, Nandha College of Pharmacy, Erode, Tamil Nadu, India.

Abstract

Transport organisations will need to incorporate the Big Data Revolution into automated vehicle systems in order to provide a safe and reliable service in inclement weather. The primary goal of this study is to see if multisource data approaches can be outperformed by deep learning models. The goal of this work is to compare the accuracy of the advanced DNN-LSTM in predicting flow time series. In addition, traffic detectors may have both temporal and spatial properties. Following training with local arterial traffic and meteorological data, traffic flow characteristics can be learned under varied precipitation situations. Time series predictions can be improved by using memory blocks containing memory cells to hold long-term and short-term properties.

Keywords: Prediction, Rainfall, Deep Learning, LSTM.

1. Introduction

When it comes to human life, rainfall is the most important factor in any form of meteorological occurrence. Human civilisation is profoundly affected by rainfall. Climate plumage is a natural, predictive, and challenging climate phenomenon. Preparation and management of water supplies, as well as reservoir service and flood control, all depend on accurate rainfall data. Precipitation also has a significant impact on urban transportation and other human activities [1]-[5].

Rainfall is particularly challenging to predict because of the large temporal and spatial fluctuations in precipitation caused by the complexities of the atmospheric processes that lead to it. As a result, accurate rainfall estimates continue to be a substantial challenge in operational hydrology, despite improvements in recent years. Plants are flourishing and life is expanding. Agriculture, which generates a large portion of the country's income, is intimately tied to precipitation [7]-[13].

The terms "knowledge discovery" (also known as "KDD") and "data mining" are frequently used interchangeably in the context of databases. Using KDD, you can elevate low-level data to high-level knowledge [6]. As a result, the term "KDD" refers to a nontrivial yet undiscovered database extraction method for implicit databases. However, in the real world of data mining, data extraction and KDD are critical steps in the KDD process. Figure 1 depicts the data collection in an iterative discovery process.

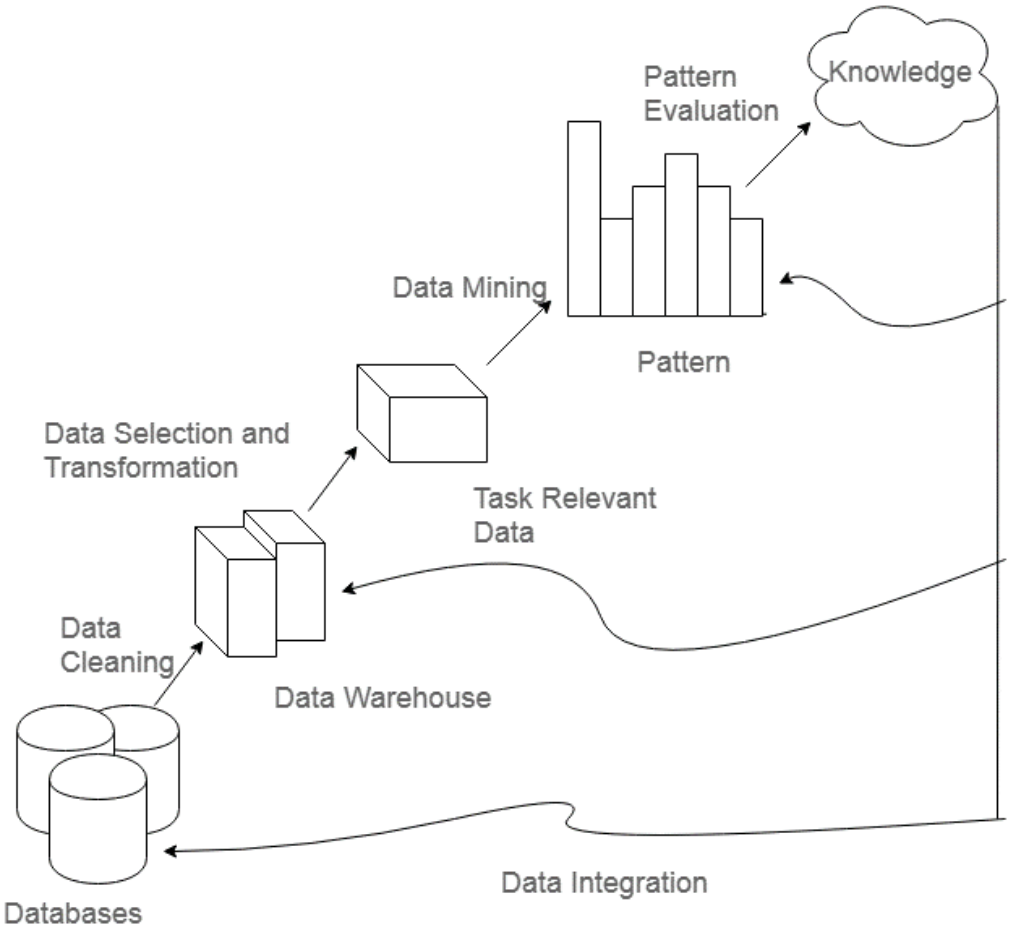

Figure 1 Knowledge Discovery Process

When it comes to determining if the models can outperform conventional methods, this study is the most important. The goal of this research is to examine the accuracy of advanced DNN-LSTM and BPNN in predicting flow time series. In addition, traffic detectors may have both temporal and spatial properties. Following training with local arterial traffic and meteorological data, traffic flow characteristics can be learned under varied precipitation situations. The results and approach can help with traffic modelling, operations, and management performance. The results can be considerable and serve as a basis for further examination of other environmental elements if the prediction accuracy is examined without additional traffic data.

2. Problem Definition

Machine learning has been shown to increase the accuracy of individual classifiers by merging several, but none of these methods resolves the issue of class imbalance. This question is specifically addressed by the learning algorithms. As a result of this examination, numerous flaws in the multiclass classification system were discovered and discussed. It's easy to see where they're coming from. Problems with multi classifier performance and correct rating, as well as the creation of unclassified zones, plague these systems. Classification accuracy and performance suffer as the unclassified area expands.

Because one class is significantly smaller than the other, there is a disparity in the number of students in each class. Under-sampling is well-known in this area. Due to the fact that it simply makes use of a majority class subdivision, under-sampling is particularly effective. The sampling method has a drawback in that it removes a lot of useful examples. In order to get around this limitation, unchecked learning methods are being developed for use in supervised education.

It was thought that class overlap was one of the most difficult classification problems to solve. When there is both class overlap and an imbalance, the problem becomes much more complicated. There are five popular categorization systems and three overlap class modelling approaches employed in the comparative study.

Learning machines have a hard time correctly classifying objects into several categories. A prescription label is made up of a number of instances of the learning set for a multi-class classification. Voting and predicting correctly in a multi-class classification scenario for imbalanced datasets is essential. Multiclass classification performance and accuracy are dependent on the prediction and voting of new class data[18]. Multiclass data Confusion is generated, performance deteriorates, and classification accuracy decreases as new categories of imbalanced data are assigned.

In this manner, the original unbalanced data samples are added. Over-sampling can be accomplished in two ways: randomly or synthetically. In the random over-sampling method, the minority samples are randomly reproduced, which can result in a fitting difficulty. Minority samples are over-sampled to create synthetic samples.

Classification is difficult when the data is skewed and the classifications overlap. For the most part, real-world data are ignored when classifying data that appears to be skewed. Using conventional approaches, the class with the most samples has been better classified than other classes. Unbalanced data sets can be classified using a variety of methods, each of which has advantages and disadvantages. The authors have described a new hierarchical decomposition method for unbalanced data sets that varies from earlier solutions to class imbalances.

3. Proposed Method

Distancing them from each other is the best option at this point. This is the next step in the process, which is to combine each piece of data into the nearest centre. The first phase will be finished when there are no outstanding issues and the group is of an appropriate age. It is therefore necessary to recalculate each cluster's new centre. Once the new centres have been formed, the identical data sets must be linked together. Step by step, the centres at k change until no further adjustments have been made or the centres have stopped.

Using K-clustering, data is grouped into K clusters and the centre of each cluster is defined. For each group, it takes several steps until the total of the squared errors becomes infeasible. The next steps are as follows: Calculate the average value of each cluster, assign each point to the cluster with the closest average value, and then measure the distance between the average value of individual cluster points.

At first, randomly select $K(\mu_k)$ centroids from the pixel data points $\{x^{(1)}, x^{(2)}, x^{(3)},\ldots, x(N)\}$, where x represents the pixel values and N refers to the number of pixels in the image. Calculate the distance $\sum\|x^i - \mu_k\|^2$ between each data point $(x(i))$ where$i = 1,2,3,\ldots,N$ and centroid(μ_k).

Attribute a group with pixel data points that is a minimum distance from the centre. Centroids update recalculation. The new centres are the mean of the data points obtained from each cluster for each cluster. Take these two steps until the centre changes.

In this step, the average value for each cluster is determined and the distance from the mean of each point in the respective cluster is determined.

$$J = \sum_{i=1}^{N}\sum_{k=1}^{K} w_{ik}\left\|x^i - \mu_k\right\|^2 \quad (1)$$

If pixel data point x_i belongs to cluster k, then w_{ik}=1; otherwise w_{ik} = 0. μ_k is the centroid of x_i cluster.

The preceding function Eq.(3.1) is the reduced equation of two parts. At first, we differentiate J with respect to w_{ik} and keep μ_k as fixed which updates cluster assignments (E-step). Then we differentiate J with respect to μ_k and keep w_{ik} fixed and recalculate the centroids after the cluster assignments from previous step (M-step). Therefore E-step is:

$$\frac{\partial J}{\partial r} = \sum_{i=1}^{N}\sum_{k=1}^{K}\left\|x^i - \mu_k\right\|^2 \Rightarrow w_{ik} = \begin{cases} 1 & if\ k = \arg\min\left\|x^i - \mu_k\right\|^2 \\ 0 & otherwise \end{cases} \quad (2)$$

In other words, assign the data points x_i of a pixel to the closest cluster, obtained by calculating the square distance sum from the centroid of the cluster.

The *M*-step is:

$$\frac{\partial J}{\partial \mu_k} = 2\sum_{i=1}^{N} w_{ik}\left(x^i - \mu_k\right) = 0$$

$$\Rightarrow \mu_k = \frac{\sum_{i=1}^{N} w_{ik}x^i = 0}{\sum_{i=1}^{N} w_{ik} = 0} \quad (3)$$

4. Results and Discussions

A Ten-fold cross validation is conducted between the training and testing datasets and the tests are conducted with various feature learning methods like K-Nearest Neighbor, Wavelet Transform, Cluster Analysis, Multiplicative Decomposition, Support Vector Machine and Seasonal Decomposition.

The method is tested under rainfall datasets collected in India between 1901 and 2015 (). The validation is segmented between 1901 – 1920, 1921 – 1940, 1941 – 1960, 1961 – 1980, 1981 – 2000 and finally between 2001 and 2020. The datasets is included with rainfall data of various state across 12 months as in Table 1.

Table 1: Datasets used for prediction

Districts	**Coefficient of variation (Annual)**	**Mean (annual)**	**Standard deviation (Annual)**
Andaman & NicobarIslands	13.6272	2916.822	396.6918
Arunachal Pradesh	31.1622	3493.072	1085.166
Assam & Meghalaya	11.4228	2604.599	295.6902
Bihar	16.2324	1207.811	195.1896
Chhattisgarh	15.531	1377.75	214.1274
Coastal Andhra Pradesh	18.2364	1055.707	191.6826
Coastal Karnataka	14.2284	3403.393	481.962

East Madhya Pradesh	18.1362	1210.216	219.1374
East Rajasthan	25.9518	653.0034	168.9372
East Uttar Pradesh	20.1402	991.7796	199.2978
Gangetic west Bengal	15.531	1492.98	230.8608
Gujarat region	30.561	924.9462	281.7624
Haryana Delhi & Chandigarh	26.6532	536.571	142.8852
Himachal Pradesh	19.7394	1270.235	249.8988
Jammu & Kashmir	20.541	1142.581	234.0672
Jharkhand	15.2304	1317.53	199.7988
Kerala	14.4288	2936.261	423.4452
Konkan & Goa	16.1322	2980.75	479.457
Lakshadweep	17.1342	1608.711	275.55
Madhya Maharashtra	18.1362	885.768	160.1196
Matathwada	23.6472	799.4958	188.4762
Naga Mani Mizo Tripura	16.9338	2468.327	416.9322
North interior Karnataka	18.8376	722.1414	135.5706
Orissa	12.9258	1462.519	188.9772
Punjab	27.7554	597.3924	165.33
Rayalseema	19.6392	767.9328	150.3
Saurashtra & Kutch	41.1822	495.2886	203.8068
South interior Karnataka	14.7294	1040.878	152.7048
Sub Himalayan West Bengal & Sikkim	12.2244	2771.031	339.4776
Tamil Nadu	17.3346	948.2928	164.4282
Telangana	21.543	955.8078	205.41
Uttarakhand	18.1362	1471.036	266.7324
Vidarbha	18.537	1098.092	202.7046
West Madhya Pradesh	19.4388	937.0704	181.4622
West Rajasthan	37.8756	289.2774	109.218
West Uttar Pradesh	22.1442	837.5718	184.869

5. Conclusions

The methods and outcomes can help enhance modelling and operational performance. Non-transport data can be used to examine the accuracy of the predictions and can serve as a reference for further considerations. Thus, it can be observed that the LSTM model combines memory blocks with memory cells to retain long and short-time features that are more suited for time series prediction.

References

[1] Sengupta, A., Jin, F., & Cao, S. (2019, July). A Dnn-LSTM based target tracking approach using mmWave radar and camera sensor fusion. In *2019 IEEE National Aerospace and Electronics Conference (NAECON)* (pp. 688-693). IEEE.

[2] Nakayama, S., & Arai, S. (2018, August). DNN-LSTM-CRF Model for Automatic Audio Chord Recognition. In *Proceedings of the International Conference*

on Pattern Recognition and Artificial Intelligence (pp. 82-88).
[3] Ozcan, A., Catal, C., Donmez, E., &Senturk, B. (2021). A hybrid DNN–LSTM model for detecting phishing URLs. *Neural Computing and Applications*, 1-17.
[4] Xueying, Z. H. A. N. G., Puhua, N. I. U., & Fan, G. A. O. (2018). DNN-LSTM based VAD algorithm. *Journal of Tsinghua University (Science and Technology)*, *58*(5), 509-515.
[5] Shah, D., Campbell, W., &Zulkernine, F. H. (2018, December). A comparative study of LSTM and DNN for stock market forecasting. In *2018 IEEE International Conference on Big Data (Big Data)* (pp. 4148-4155). IEEE.
[6] He, T., &Droppo, J. (2016, March). Exploiting LSTM structure in deep neural networks for speech recognition. In *2016 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)* (pp. 5445-5449). IEEE.
[7] Khan, S. A., & Chang, H. T. (2019). Comparative analysis on Facebook post interaction using DNN, ELM and LSTM. *PloS one*, *14*(11), e0224452.
[8] Żołna, K., &Romański, B. (2017, February). User modeling using LSTM networks. In *Thirty-First AAAI Conference on Artificial Intelligence*.
[9] Li, Y., Ghosh, S., Joshi, J., &Oviatt, S. (2020, November). Lstm-dnn based approach for pain intensity and protective behaviour prediction. In *2020 15th IEEE International Conference on Automatic Face and Gesture Recognition (FG 2020)* (pp. 819-823). IEEE.
[10] S Veeramani, S Karuppusamy,"A survey on sentiment analysis technique in web opinion mining" International Journal of Science and Research (IJSR) – Volume 3 Issue 8– 2012
[11] S. Karuppasamy2 S. Gowtham1," A Study on Data Mining Information Security" International Journal of Research and Advanced Development (IJRAD), ISSN: 2581-4451 Volume 3, Issue 02- June 2019
[12] S. Karuppusamy and G. Singaravel , "Investigation Analysis for Software Fault Prediction using Error Probabilities and Integral Methods" Applied Mathematics & Information Sciences An International JournalVolume 13 Issue S1- 2019
[13] Thirumoorthy, P., Kalyanasundaram, P., Karupusamy, S., &Prabhu.S, "ENERGY EFFICIENT ROUTING IN WSNS USING DELAY AWARE DYNAMIC ROUTING PROTOCOL" Journal of Critical Reviews, 6(6), 455-459. https://doi.org/10.31838/jcr.06.06.70
[14] Tan, T., Qian, Y., Yu, D., Kundu, S., Lu, L., Sim, K. C., ... & Zhang, Y. (2016, March). Speaker-aware training of LSTM-RNNs for acoustic modelling. In 2016 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP) (pp. 5280-5284). IEEE.
[15] Zhen, T., Yan, L., & Yuan, P. (2019). Walking gait phase detection based on acceleration signals using LSTM-DNN Algorithm. Algorithms, 12(12), 253.
[16] Hlaing, A. M., Pa, W. P., & Thu, Y. K. (2019). Enhancing Myanmar speech synthesis with linguistic information and LSTM-RNN. In Proc. 10th ISCA Speech Synthesis Workshop (pp. 189-193).
[17] Song, E., Soong, F. K., & Kang, H. G. (2017). Effective spectral and excitation modelling techniques for LSTM-RNN-based speech synthesis systems. IEEE/ACM Transactions on Audio, Speech, and Language Processing, 25(11), 2152-2161.
[18] Manchula A., Arumugam S., "Face and fingerprint biometric fusion: Multimodal feature template matching algorithm", International Journal of Applied Engineering Research, Volume 9, Issue 22, Pages 17295-17315, 2014