IJITCE

ISSN 2347- 3657

International Journal of Information Technology & Computer Engineering

Email : ijitce.editor@gmail.com or editor@ijitce.com

International Journal of Information Technology & Computer Engineering

ISSN 2347–3657

Volume 12, Issue 3, 2024

A Machine Learning Approach for Rainfall Estimation Integrating Heterogeneous Data Sources

[1] K SUPARNA, [2]R. DEEPIKA

[1](Assistant Professor), MCA, **DANTULURI NARAYANA RAJU COLLEGE(A) PG COURSES, BHIMAVARAM ANDHRA PRADESH**

[2]MCA, scholar, **DANTULURI NARAYANA RAJU COLLEGE(A) PG COURSES, BHIMAVARAM ANDHRA PRADESH**

ABSTRACT

Providing an accurate rainfall estimate at individual points is a challenging problem in order to reduce risks derived from several rainfall events, such as floods and landslides. Dense networks of sensors, named rain gauges(RGs), are typically used to obtain direct measurements of precipitation intensity in these points. These measurements are usually interpolated by using spatial interpolation methods for estimating the precipitation field over the entire area of interest. However, these methods are computationally expensive, and to improve the estimation of the variable of interest in unknown points, it is necessary to integrate further information. To overcome these issues, this work proposes a machine learning-based methodology that exploits a classifier based on ensemble methods for rainfall estimation and is able to integrate information from different remote sensing measurements. The proposed approach supplies an accurate estimate of the rainfall where RGs are not available, permits the integration of heterogeneous data sources exploiting both the high quantitative precision of RGs and the spatial pattern recognition ensured by radars and satellites, and is computationally less expensive than the interpolation methods. Experimental results, conducted on real data concerning an Italian region, Calabria, show a significant improvement in comparison with Kriging with external drift (KED), a well-recognized method in the field of rainfall estimation, both in terms of the probability of detection (0.58 versus 0.48) and mean-square error (0.11 versus 0.15).

ISSN 2347–3657

Volume 12, Issue 3, 2024

1.INTRODUCTION

Accurate rainfall estimate is crucial for flood hazards protection, river basins management, erosion modeling, and other applications for hydrological impact modeling. To this aim, rain gauges (RGs) are used to obtain a direct measurement of intensity and duration of precipitations at individual sites.

In order to estimate rainfall events in areas not covered by RGs, interpolation methods computed on the basis of the values recorded by these RGs are used. Many variants of these methods have been proposed in the literature, and among them, the Kriging geo statistical method is one of the most used and recognized in the field.

An accurate spatial reconstruction of the rainfall field is a critical issue when dealing with heavy convective meteorological events. In particular, convective precipitations can produce highly localized heavy precipitation, not detected by sparse RGs, and floods can arise without a rainfall being detected. To overcome this issue, a recent trend in the literature is to integrate heterogeneous rainfall data sources to obtain a more accurate estimate by using interpolation methods.

Unfortunately, the largely used ordinary Kriging (OK) can exploit only one source of data as input; therefore, Kriging with external drift (KED) was one of the most popular approaches adopted to overcome this limitation. Indeed, KED allows a random field to be interpolated, and different from the OK, it is able to take into account secondary information. The main problem is that these methods are computationally expensive and require a large number of resources to work properly.

A different approach relies on exploiting machine learning (ML) techniques. However, using these methods requires coping with different hard issues, i.e., unbalancing of the classes, a large number of missing attributes, and the need for working incrementally as soon as new data are available. Typically, ensemble methods are used to address these issues. Ensemble is a classification technique, in which several models, first trained by using different classification algorithms or samples of data, are then combined to classify new unseen instances. In comparison with the case of using a single classification model, the ensemble paradigm permits handling the problem of unbalanced classes and reducing the variance and the bias of the error.

ISSN 2347–3657

Volume 12, Issue 3, 2024

Especially, ensemble-based techniques can be used to address the issues concerning the rainfall estimation and to support the monitoring of meteorological (intense) events. These methods are also able to capture nonlinear correlations (e.g., relations between sensor data, cloud properties, and rainfall estimate). In order to address the main issues of rainfall estimation, in this article, an ML-based methodology, adopting a hierarchical probabilistic ensemble classifier (HPEC) for rainfall estimation, is introduced. The proposed approach, by integrating data coming from different sources (i.e., RGs, radars, and satellites) and exploiting an under-sampling technique for handling the unbalanced classes problem typical of this scenario, permits accurate estimation of the rainfall where RGs are not available.

Our approach is an effective solution for real scenarios, as in the case of an officer of the Department of Civil Protection (DCP), who has to analyze the rainfall in a specific zone presenting risks of landslides or floods. The experimental

evaluation is conducted on real data concerning Calabria, a region located in the South of Italy, and provided by the DCP. Calabria is an effective test ground because of its stro ng climate variability and its complex orography.

Our contributions can be summarized as follows.

1) Three heterogeneous data sources (i.e., RGs, radar, and Metaset) are integrated to generate more accurate estimates of rainfall events.

2) Different classification methods are compared on a real case concerning Calabria, a southern region in Italy, and a hierarchical probabilistic ensemble approach is proposed.

3) Different ML-based methods, pre trained only on historical data, with a widely used interpolation method in the hydrological field (i.e., KED) are compared.

The rest of this article is organized as follows. In Section II, some related works are analyzed, and the main differences with our approach are noted. Section III illustrates the case study and describes the main sources of data used by the framework. In Section IV, the methodology used to estimate the rainfall is specified. Section V is devoted to some experimental results and discussion.

ISSN 2347–3657

Volume 12, Issue 3, 2024

2. EXISTING SYSTEM

An existing system is based on the ensemble paradigm include the work in which, similar to our work, employs a probabilistic ensemble and merges two sources of data (i.e., rain gauges and radar) even if the aim of this work is to develop a run-off analysis. Afterward, a blending technique is applied to the results of the runoff hydrologic models to determine a single runoff hydrograph. Experimental results show that the hydrologic models are accurate and can help to make more effective decisions in the flood warning. Frei and Isotta define a technique for deriving a probabilistic spatial analysis of daily precipitation from rain gauges. The final model represents an ensemble of possible fields, conditional on the observations, which can be explained as a Bayesian predictive distribution measuring the uncertainty due to the data sampling from the station network. An evaluation of a real case study, located in the European Alps, proves the capability of the approach in providing accurate predictions for a hydrological partitioning of the region.

The work in proposes an interesting study of the daily precipitations for Australia and several regions of South and East Asia, based only on high-resolution gauges. Basically, the adopted model can be figured out as a mean of the analyses generated for each source. The authors highlight how the ensemble approach outperforms the single members composing the model in terms of global accuracy. Moreover, the proposed model is also able to capture additional information from different precipitation products. Both the last two works exploit an ensemble scheme to provide more accurate predictions, proving the capability of ensemble methods to ensure good results also in a rainfall estimate scenario. However, different from our work, the adopted combination strategies are quite simple, and a combination of heterogeneous data sources is not considered.

Calabria, Chiaravalloti *et al* studied the performance of three recently developed satellite-based products, i.e., IMERG, SM2RASC, and a clever combination of SM2RASC and IMERG using as a benchmark both RG only data and the integrated RG-radar product. Experiments permit to establish that IMERG has good performance at time resolutions higher than 6 h, and the combination of IMERG and SM2RASC obtains a higher quality satellite rainfall product. Most of the other

ISSN 2347–3657

Volume 12, Issue 3, 2024

approaches integrate data from different sources, i.e., satellite channels and radars. Some of them are based on the identification of suitable models that exploit the relation between optical and microphysical properties of clouds and use the data to find the appropriate parameters for these models. Other works individuate the models by using statistical techniques. For instance, Bayesian estimation is used in order to provide precipitation estimations based on satellite multispectral data; reference estimates are provided by methods that use radar data as input.

Verdin *et al* also adopt Bayesian estimation in order to estimate the parameters of the model; their system integrates RG observations and satellite data and adopts an interpolation technique based on the Kriging method. All these techniques are able to provide interesting results, but they require a rather delicate phase of parameters estimation of the particular model; therefore, as a side effect, usually, their flexibility and effectiveness tend to be hampered. As the relations between sensors data, cloud properties, and rainfall estimates are highly nonlinear, more flexible approaches based on ML techniques have been investigated recently. For instance, the problem of detecting convective events and closely related rainy areas is addressed in by using ANNs combined with support vector machines. Data sets are obtained by processing data coming from optical channels of the multispectral instrument onboard of Metaset Second Generation (MSG) satellites; different from our work, RG measures are used only as a reference but not in the training phase of the algorithm. Sehat et alpropose an approach to rainfall estimation based on SVMs; the input data are integrated from multispectral channels on MSG; and two models are developed for daytime and nighttime respectively.

Results are compared to similar approaches based on ANNs, and random forest (RF) and RGs are used only to validate the approach. Another approach based on ANNs is described in this work, given as an image matrix, radar data are used as reference in detecting rainy pixels. Kuhnian *et al* also adopt the ensemble paradigm and, in particular, employ RFs to infer rainfall rates from data coming from multispectral channels on MSG satellites.

Disadvantages

ISSN 2347–3657

Volume 12, Issue 3, 2024

- The system is not implemented hierarchical probabilistic ensemble classifier (HPEC) for rainfall prediction.
- The system is implemented artificial neural networks (ANNs) as a forecasting method in which prediction is not accurate.

3. PROPOSED SYSTEM

Our approach is an effective solution for real scenarios, as in the case of an officer of the Department of Civil Protection (DCP), who has to analyze the rainfall in a specific zone presenting risks of landslides or floods. The experimental evaluation is conducted on real data concerning Calabria, a region located in the South of Italy, and provided by the DCP. Calabria is an effective test ground because of its strong climate variability and its complex orography. Our contributions can be summarized as follows.

1) Three heterogeneous data sources (i.e., RGs, radar, and Metaset) are integrated to generate more accurate estimates of rainfall events.

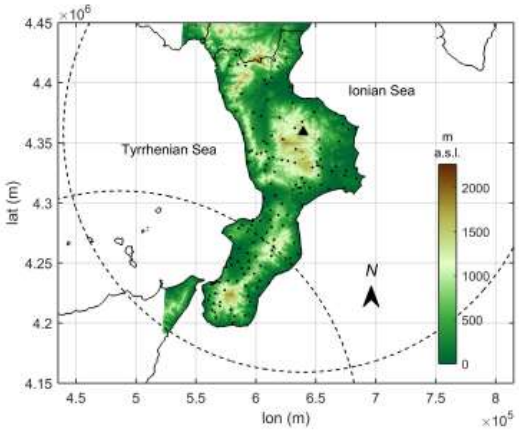
2) Different classification methods are compared on a real case concerning Calabria, a southern region in Italy, and a hierarchical probabilistic ensemble approach is proposed.

3) Different ML-based methods, pre trained only on historical data, with a widely used interpolation method in the hydrological field (i.e., KED) are compared.

Advantages

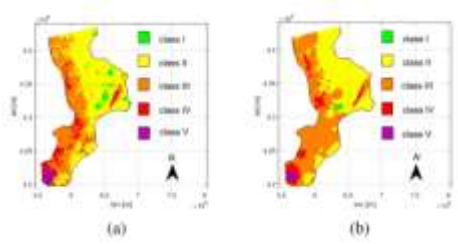
- In the proposed system, raw data are preprocessed to make them suitable for the analysis, and an under-sampling strategy is adopted to address the class unbalanced problem.
- The proposed system developed an Effect of Integrating RG, Satellite, and Radar Measurements and are tested and trained with an effective ML Classifiers.

4.OUTPUTSCREENS



ISSN 2347–3657

Volume 12, Issue 3, 2024



(a) (b)

5.CONCLUSION

An ML-based approach for the spatial rainfall field estimation has been defined. By integrating heterogeneous data sources, such as RGs, radars, and satellites, this methodology permits estimation of the rainfall, where RGs are not present, also exploiting the spatial pattern recognition ensured by radars and satellites. After a phase of preprocessing, a random uniform under sampling strategy is adopted, and finally, an HPEC permits the model used to be built to estimate the severity of the rainfall events. This ensemble is based on two levels: in the first level, a set of RF classifiers are trained, while, in the second level, a probabilistic metal earner is used to combine the estimated probabilities provided by the base classifiers according to a stacking schema. Experimental results conducted on real data provided by the Department of Civil Protection show significant improvements in comparison with Kriging with external drift, a largely used and well-recognized method in the field of rainfall estimation. In particular, the ensemble method exhibits a better capacity in detecting the rainfall events. Indeed, both the POD (0.58) and the MSE (0.11) measures obtained by HPEC are significantly better than the values obtained by KED (0.48 and 0.15, respectively). As for the last two classes, representing intense rainfall events, the difference between the Kriging method and HPEC is not significant (in terms of F-measure) although HPEC is computationally more efficient.

Indeed, the complexity of the Kriging method is cubic in the number of the samples [51], which makes the procedure really expensive from the computational point of view, when a large number of points are analyzed. On the contrary, the ML algorithms (i.e., RF) exhibit a quadratic complexity. Moreover,

ensemble methods are highly scalable and parallelizable. Therefore, we believe that our approach has some relevant advantages in this field of application.

In addition, by analyzing the effect of the integration of the different sources of data, it is evident that all the data sources contribute to the good performance of the technique. In particular, by removing the RG information,

ISSN 2347–3657

Volume 12, Issue 3, 2024

the performance of the algorithm worsens the sensibly for all the measures. In the cases of the removal of one of the other two types of data, the degradation is less evident; however, the lowest value (0.11) of the MSE is obtained when all the data are used, which confirms that it is necessary to use all the sources of data to obtain better results.

As future work, we plan to validate the method on a larger time interval, in order to consider effects due to seasonal and yearly variability, also considering the possibility of incrementally building the flexible ensemble model with the new data.

6. REFERENCES

1.J. E. Ball and K. C. Luk, "Modeling spatial variability of rainfall over attachment," J. Hydrologic Eng., vol. 3, no. 2, pp. 122–130, Apr. 1998.

2.S. Ly, C. Charles, and A. Degree, "Different methods for spatial interpolation of rainfall data for operational hydrology and hydrological modeling at watershed scale. a review," Biotechnologies, Agronomie, Société et Environment, vol. 17, no. 2, p. 392, 2013.

3.H. S. Wheater et al., "Spatial-temporal rainfall fields: Modelling anstatistical aspects," Hydral. Earth Syst. Sci., vol. 4, no. 4, pp. 581–601, Dec. 2000.

4.J. L. McKee and A. D. Binns, "A review of gauge–radar mergingmethods for quantitative precipitation estimation in hydrology," Can.Water Resour. J./Revue Canadienne des Ressources Hydriques, vol. 41, nos. 1–2, pp. 186–203, 2016.

5.F. Cecinati, O. Wani, and M. A. Rico-Ramirez, "Comparing approaches to deal with non-gaussianity of rainfall data in Kriging-based radar gauge rainfall merging," Water Resoure. Res., vol. 53, no. 11, pp. 8999–9018, Nov. 2017.

6.H. Wackernagel, Multivariate Geo statistics: An Introduction with Applications. Berlin, Germany: Springer, 2003.

7.L. Breiman, "Bagging predictors," Mach. Learn., vol. 24, no. 2, pp. 123–140, Aug. 1996.

8.B. J. E. Schroeter, Artificial Neural Networks in Precipitation Now-Casting: An Australian Case Study. Cham, Switzerland: Springer, 2016, pp. 325–339.

8.X. Shi, Z. Chen, H. Wang, D. Yeung, W. Wong, and W. Woo, "Convolutional LSTM network: A machine learning approach for precipitation now casting," in Proc. 28th Int. Conf. Neural Inf.

ISSN 2347–3657

Volume 12, Issue 3, 2024

Process. Syst., vol. 1, Dec. 2015, pp. 802–810.

9.W.-C. Hong, “Rainfall forecasting by technological machine learning models,” Appl. Math. Comput., vol. 200, no. 1, pp. 41–57, Jun. 2008.

10.A. Parmar, K. Mistree, and M. Sompura, “Machine learning techniques for rainfall prediction: A review,” in Proc. 4th Int. Conf. Innov. Inf.Embedded Common. Syst. (ICIIECS), Mar. 2017, pp. 152–162.