Smart sensing-based approaches for Real-time River Water Pollution Monitoring

वास्तविक समय नदी जल प्रदूषण निगरानी के लिए स्मार्ट सेंसिंग-आधारित दृष्टिकोण

Thesis submitted in partial fulfillment for the Award of Degree

Doctor of Philosophy

by

Swati Sandeep Chopade

स्वाति संदीप चोपड़े

DEPARTMENT OF COMPUTER SCIENCE AND ENGINEERING
INDIAN INSTITUTE OF TECHNOLOGY
(BANARAS HINDU UNIVERSITY)
VARANASI - 221005

Roll No. 18071015

Year 2024

Chapter 7

Conclusion and future work

This chapter recalls the context of the thesis, summarizes the main contributions of this work, and finally outlines the future research directions towards river water pollution monitoring maximizing the network lifetime.

7.1 Conclusion

This thesis has studied various problems encountered for identifying river water pollution level. The principal objective of this work was to develop effective and efficient river water pollution assessment and monitoring system using deep neural network. This work considered two tasks, *i.e.*, river water pollution level assessment system using deep neural network and energy-efficient river water pollution monitoring system using game theory. The decision to use deep neural networks for sensor-based river water pollution assessment and game theory for monitoring via knowledge distillation was carefully evaluated. Through extensive experiments, we found that deep neural networks effectively captured the complex patterns in the pollution data, significantly enhancing prediction accuracy compared to traditional methods. Similarly, applying game theory provided a strong framework for analyzing interactions among stakeholders, leading to more effective monitoring strategies. Our results showed that both

approaches not only justified their selection but also offered valuable insights and improvements in the assessment and monitoring processes. While alternative methods exist, the findings indicated that the chosen approaches were well-suited to the specific challenges of our application. Therefore, we believe that these methodologies were appropriate and beneficial to the objectives of our study.

During assessing river water pollution level, we had unlabelled limited lab dataset and noisy labels in the training dataset due to the automatic annotation. Furthermore, we considered the problems of the resources constraints (processing power, storage, and battery) of IoT devices during deployment of deep neural network on resources limited devices to estimate river water pollution level. Moreover, we considered the problems of energy consumption in transmitting the water pollution data to the base station. These problems are more common and frequently occur in real-world scenarios; therefore, employing suitable mechanisms results in the effective assessment and monitoring of river water pollution. This thesis has made four major research contributions to resolve the problems encountered while assessing and monitoring river water pollution.

In chapter 3, we covered the problem of river water pollution level identification using unlabelled limited lab data collected using the instruments available in the lab and automatic annotation of limited lab data. We proposed a mathematics-based model to estimate WQI and automatically label the limited lab dataset. The mathematical model constructed by selecting the distinct water parameters to estimate WQI. After selection of distinguishable parameters, established weights for those parameters. We employed a weight assignment technique for various water parameters, incorporating the weights established by the NSF. Assigning equal weights to all parameters would be impractical, as each parameter has a different impact on assessing water quality. Therefore, we assigned unequal weights to the six selected parameters, basing the calculations on NSF guidelines. Further, we calculated sub-indices for those selected parameters. Finally, aggregated the weights and sub-indices to get final WQI. Next, we automatically

assigned the labels to limited lab dataset.

In Chapter 4, we addressed the second challenge: the automatic annotation of sensory data collected from the Hanna multi-parameter sensor and the assessment of river water pollution, focusing on the issue of noisy labels in the dataset due to automatic annotation. Unlike existing methods, our model relied on GPS coordinates to automatically annotate the sensory data. The main motivation for using GPS coordinates is to precisely tag the location where data is collected, as water behavior varies at different points due to factors such as industrial waste and city drains flowing into the river. Identifying the location and its contributing factors is crucial during data collection, making GPS essential. Moreover, collecting GPS coordinates separately from the data would require specialized tools, significantly increasing the overall cost of estimation. We introduced an automatic label transfer mechanism to streamline the annotation process. The chapter presented several experiments to evaluate the performance of the proposed model on the river water dataset. To tackle the noisy label problem, we proposed a deep learning-based approach for recognizing river water pollution levels using large volumes of sensory data with noisy labels. Unlike previous approaches, this deep neural network model built a classifier to improve recognition without prior information about the extent of noisy labels. We are working with time-series river water data, making LSTM deep neural networks an ideal choice for classifying water pollution data. LSTMs are specifically designed to handle sequential data like time series by capturing temporal dependencies over long durations. The noise-handling mechanism incorporated a specialized loss function that minimized the gap between true and predicted labels using a dynamic variable. We estimate the predicted probability during the training of the model. Extensive experiments were conducted to validate the effectiveness of the approach using the collected river water dataset.

In Chapter 5, we proposed a system to address the resource constraints of IoT devices during the river water pollution identification task. This chapter introduced a

knowledge distillation approach to estimate pollution levels using IoT devices. Unlike existing methods, our approach involved constructing and training a cumbersome model first. We then applied filter-level pruning to compress the large model, making it suitable for deployment on IoT devices. The compressed DNN was further trained using a knowledge distillation approach, with the cumbersome model serving as a guide. The compressed DNN was successfully deployed on resource-limited IoT devices, achieving acceptable accuracy in detecting river water pollution. Additionally, we conducted extensive experiments to validate the effectiveness of the proposed approach using a collected river water dataset.

In Chapter 6, we introduced an energy-efficient river water pollution monitoring system that utilizes deep neural networks and long-range communication technology. Unlike existing approaches, we proposed a game theory-based method to transmit pollution data from the river to the base station while minimizing energy consumption. Sensory data is considered crucial if the difference between the values of its attributes and those of existing data samples at a given location exceeds a defined threshold. Typically, crucial data is captured by sensors immediately after a change in the Water Quality Index (WQI) class. To establish a threshold for identifying crucial sensory data, we begin by focusing on key attributes (e.g., temperature, pH, DO) and collecting both new and existing sensory data. We then analyze the distribution of the existing data to pinpoint where significant changes in the Water Quality Index (WQI) occur. At these change points, we calculate the differences between the new and existing data samples and set a threshold using the 95^{th} percentile statistical method. This threshold is validated against a separate dataset and adjusted as necessary to optimize sensitivity and specificity, with continuous monitoring as new data is gathered. Since parameters like pH, dissolved oxygen (DO), and electrical conductivity are used, the threshold value varies depending on the specific data collection location along the river. This threshold is location-specific and varies according to GPS coordinates, and we have

already mapped the relationship between each location and its corresponding threshold value. This approach calculates the optimal time duration for selecting the appropriate spreading factor to efficiently transmit data from the river to the central remote host. By doing so, it reduces energy usage and ensures reliable data transmission. Additionally, we conducted a real-world study to assess the feasibility and performance of the proposed technique.

7.2 Future Research Directions

This thesis work has made contributions towards the investigation of the different challenges encountered while identifying the river water pollution level. The primary objective is to mitigate the negative impact of noisy labels in the river water dataset introduced due to automatic annotation. Further, we introduced different mechanisms to successfully assess and monitor the river water pollution. We obtained the acceptable accuracy in detecting the water pollution level despite limited resources of IoT devices. By using the results in our work, one can effectively design appropriate approaches to handle different challenges encountered while assessing and monitoring the river water pollution level. We believe that the proposed work would motivate further research towards different IoT applications such as healthcare, air quality assessment and sound noise monitoring. The work described here can be extended further in the following research directions:

- This work also provides a further research direction towards incorporating the deep learning and automated annotation technique during unsupervised learning.
- This work also provides a future direction to develop the LN deployment strategy that reduces the number of LNs and prolong the network lifetime.
- This work only covered the problem of noisy labels in the training datasets; however, the persistence of noise in the sensory instance along with noisy labels is still a research challenge. Thus, handling noise in both instances and labels could

be the future work in this area.

- We also plan to consider the imbalanced dataset while training a recognition model.