EXTENDED ABSTRACT

Learning Optimal Decision Criteria for Early Classification

Name of the student: **Anshul**

Roll number: **16071502**

Degree for which submitted: **Ph.D.**

Supervisor: **Prof. Sanjay Kumar Singh**

DEPARTMENT OF COMPUTER SCIENCE AND ENGINEERING

INDIAN INSTITUTE OF TECHNOLOGY

(BANARAS HINDU UNIVERSITY)

VARANASI - 221005

Early classification of time series is valuable in many real-world applications where data is generated over time. For example, if a disease is diagnosed early for a patient from a series of medical observations, it will reduce the treatment cost as well as the recovery time. Also, early diagnosis could save the patient's life by allowing a significant treatment time to the healthcare organization before the disease displays a complete effect on the patient. In agricultural monitoring, timely prediction of droughts and shortage of multiple resources would enable the implementation of necessary measures to prevent famine and determine sustainable policy. The aim of early classification is to predict the class label of incoming time series as early as possible before observing its complete sequence. In general, whenever early prediction time improves, the prediction accuracy decreases. In other words, one can achieve better accuracy by waiting for more data points in the series, but it will delay the response time. In time-sensitive applications, it is worth sacrificing some classification accuracy in favour of early predictions, preferably early enough for taking actionable decisions. Thus, there exists a trade-off between earliness and accuracy. However, existing approaches do not consider trade-off optimization well in their decision criteria.

Time Series Classification (TSC) is one of the major research areas that developed over the past few years, mainly due to its practical applicability in various domains such as agriculture, healthcare, medicine, finance, and industries. The main objective of TSC is to maximize prediction accuracy. In contrast, an early classification of time series has two conflicting objectives, i.e., accuracy and earliness. Nowadays, the early classification of time series attracts researchers more due to its useful applications in various domains such as early disease prediction, early gas leakage prediction, drought prediction, etc.

The primary focus of this thesis is to solve the early classification problem for the time series data (both univariate and multivariate) by learning optimal decision criteria that achieve the trade-off between accuracy and earliness. The problem of early classification has

been identified as the composition of two sub-problems. The first one is to design the early classifier that can label the incomplete time series. The second is to define the decision criteria that can estimate the right time for making an online decision. Initially, we propose an early classification model for Univariate Time Series (UTS), which relies on two factors (i) a set of probabilistic classifier and (ii) a confidence threshold. The decision policy of the proposed model has two key components, i.e., safeguard point and confidence threshold. The safeguard points are determined by user-defined parameters to ensure the desired accuracy. The confidence threshold is defined by measuring the uncertainty in predicted output to ensure the reliability of class prediction. In this method, decision policy is more inclined toward accuracy and does not consider trade-off optimization. In this regard, a further optimization-based approach has been adapted for early classification and defines the early stopping rules for optimal decision making. Moreover, these rules are learned through optimization between accuracy and earliness simultaneously.

Furthermore, this optimization-based approach has been extended for Multivariate Time Series (MTS), which is more challenging than UTS because of the multiple variables involved in decision making. An ensemble-based system has been designed to label the incomplete MTS, and collective output from all the variables has been utilized for decision making. These proposed methods are highly effective for small training data sets, but feature transformation is required for training the classifiers. Finally, a deep learning-based hybrid classifier has been proposed that can capture the temporal information from the raw sensory data effectively to perform the classification task. Moreover, the optimal confidence threshold has been defined by balancing the trade-off between accuracy and earliness.

The proposed early classification approaches have been evaluated on a broad range of publicly available synthetic as well as real datasets and they demonstrated a decent trade-off between accuracy and earliness. Moreover, the usefulness of the early classification

approaches have been shown in time-sensitive applications, including malware detection, fault detection, and transportation mode detection.

LIST OF PUBLICATIONS

Refereed Journal Papers

- **Anshul Sharma** and Sanjay Kumar Singh, “A novel approach for early malware detection,” *Transactions on Emerging Telecommunications Technologies*, Wiley 2020. (**IF: 1.670**)
- **Anshul Sharma** and Sanjay Kumar Singh, “Early classification of multivariate data by learning optimal decision rules,” *Multimedia Tools and Applications*, Springer Science and Business Media LLC, 2020. (**IF: 2.600**)
- **Anshul Sharma**, Sanjay Kumar Singh, S. S. Udmale, A. K. Singh and R. Singh “Early transportation mode detection using Smartphone sensing data,” *IEEE Sensors Journal*, Institute of Electrical and Electronics Engineers,1-1, 2020. (**IF: 3.780**)
- A. G. Nath, **Anshul Sharma**, S. S. Udmale, and Sanjay Kumar Singh, “An early classification approach for improving structural rotor fault diagnosis,” *IEEE Transactions on Instrumentation and Measurement*, 2021, vol. 70, pp. 1-13. (**IF:3.890**)
- **Anshul Sharma** and Sanjay Kumar Singh, "Early classification on sequence data: Overview, Challenges, and Opportunities," *IEEE Computational Intelligence*, 2021. (Under Review). (**IF: 9.083**)

Refereed Conference Papers

- **Anshul Sharma** and Sanjay K. Singh, “Early classification of time series based on uncertainty measure,” IEEE conference on information and communication technology, 2019.

Refereed Book Chapter

- **Anshul Sharma**, *et. al.*, "Time series data representation and dimensionality reduction techniques," In Algorithms for intelligent systems. Springer Singapore, 2020.