Learning Optimal Decision Criteria for Early Classification

Thesis submitted in partial fulfillment
for the Award of Degree

Doctor of Philosophy

by

Anshul

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

Roll No. 16071502

Year 2021

Chapter 6

Conclusion and Future Directions

This chapter summarizes the important conclusions obtained from the contributions in this thesis. Additionally, it provides promising future directions to explore the problem of the early classification further.

6.1 Conclusion

In this thesis, we studied the problem of early classification of time series data by learning optimal decision criteria. The objective of early classification is to predict the class label of time series as early as possible with acceptable accuracy. The early classification problem is applicable in many domains, where data points are obtained over time. Moreover, it is highly desirable, where either collecting data points are expensive or timely decision is required.

In Chapter 2, we reviewed the existing literature on early classification of time series to find the research gap and limitations of the existing works. Broadly early classification approaches can be categorized into three groups: instance-based, shapelet-based and model-based. Shapelet-based methods are highly interpretable to the user. However, they have some limitations. Firstly, they are highly computationally expensive. Secondly, it is likely very hard to define the shapelet threshold if the time series be-

long to different class groups and do not have distinguishable patterns. On the other hand, model-based approaches are computationally moderate, but they are lacking in interpretability.

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 policy that can estimate the right time for making an online decision. Basically, the early classification problem has two conflicting objectives, i.e., accuracy and earliness. Existing approaches consider that the balancing between accuracy and earliness is essential for early classification problems. Even a very few methods have considered trade-off optimization between these two objectives.

In Chapter 3, we addressed the problem of early classification on univariate time series. A series of probabilistic classifiers have been developed to predict the class label for incomplete time series. Then two different strategies have been designed for decision making. The first method has been designed based on two critical aspects safeguard point and confidence threshold. The safeguard point reduces the unnecessary overhead of training the classifiers and ensures the desired accuracy. The confidence threshold ensures reliability in class prediction defined by measuring uncertainty in the predicted output. In the proposed approach, we have analyzed the impact of different probabilistic classifiers such as Naive bays, SVM, and GP. The GP classifier provided a good approximation of class labels as compared to others.

To achieve the trade-off between accuracy and earliness is a key challenge. However, the proposed early decision criterion has not taken it into consideration and is inclined toward accuracy only. Thus, the second method considered an optimization-based approach and designed the early stopping rules that have been learned by optimizing the trade-off between accuracy and earliness. The proposed model demonstrated good balance between accuracy and earliness as compared to the other methods when evaluated

on publicly available synthetic as well as real datasets. Moreover, the applicability of the proposed approach has been validated for early malware detection on the publicly available malware API call sequence dataset and demonstrated decent performance. These two approaches have been validated on UTS problem.

The many real-world applications generate multivariate time-series data that is more challenging compared to univariate time series. Thus we have extended the optimization-based early classification approach for MTS data in Chapter 4. In the proposed method, we have developed a series of probabilistic classifiers for each variable separately to capture the variate-wise information and adopted an ensemble-based classification approach to predict the class label for incomplete time series. Moreover, ESRs have been proposed to perform early decision tasks. In the proposed method, the trade-off between accuracy and earliness has been defined through α parameter. The proposed approach has been analyzed on existing real-world datasets, and it is found that the model is not generalized. In fact, the trade-off between accuracy and earliness depends on the characteristics of application data. However, the proposed model is able to maintain a good balance between earliness and accuracy.

The above methods have two limitations in terms of defining baseline classifier. First, a series of probabilistic classifiers have been developed for labelling the incomplete time series. Moreover, the number of classifiers depends on the number of data points in a complete time series. Second, feature transformation is needed for training the classifiers.

Therefore, in Chapter 5, we have proposed an early classification approach to overcome these issues by developing a deep learning-based early classifier that can capture hidden patterns from raw sensory data directly. The proposed model adapted an imputation-based approach for labelling the incomplete time series, and decision criterion is defined as the reliability threshold. To test the effectiveness of the proposed model, we have considered the problem of early transportation mode detection based

on smartphone sensor data. The proposed model has been evaluated on two real-world transportation data sets and demonstrated excellent performance. Besides, it has been observed that the hybrid DL model is able to capture the temporal information from the raw time series more effectively compared to the individual DL models.

6.2 Future directions

Based on the research work presented in this thesis, the following are promising future directions to explore more.

- The problem of early classification has two sub-problems, (i) designing of the early classifier and (ii) developing of good decision policy. The design of decision policy is a crucial part of an early classification problem. In the future, more complex weighted ESRs can be designed by assigning the higher weight to more informative components in MTS. Furthermore, the proposed model can be optimized for specific applications such as early voice detection, and gait recognition.
- Interpretability is also a desirable parameter in many applications for making an acceptable decision for the user in field, such as health, agriculture, etc. Therefore, to tackle early classification problem, developing interpretable decision rules with trade-off optimization between accuracy and earliness can be a potential future direction.
- This work does not consider multimodal data; therefore, the development of application-specific early classifier by considering multimodal data can be a good research direction. For example, driving behaviour analysis is a potential problem in ITS that can be monitored using multimodal data such as steering wheel angle, acceleration pressure, and the gear shift position.
- Deep learning models have automatic feature extraction capabilities. Therefore, in the future, domain-specific early classification approaches can be developed by adding the capability to handle unseen class labels while making an early decision. In this line, transfer learning and federated learning could be helpful.