Chapter 1

Introduction

Classification is an important research topic in the field of data mining and machine learning. Basically, classification is a supervised learning task in which first a prediction model is trained from seen examples (training samples), and later it is used to classify the unseen samples [1]. It has numerous real-world applications such as cancer detection [2, 3], spam filtering [4], fault classification [5], and patient monitoring [6]. The classification task has been considered extensively on structural data, where training samples are represented in the form of feature vector. Usually, these features are called attributes, which can be categorical or numeric type. However, in many real-world applications, data is collected in a more complex form that may be structured, semi-structured or unstructured, such as text data, biological data, time-series data, and multimedia data. This thesis addresses the classification problem in time series data.

Time series is an ordered sequence of measurements, called data points, recorded over time. Generally, the term time series refers to the Univariate Time Series (UTS), where only one variable is involved in measurements, such as measuring the temperature of a room every second, electrical activity of a patient's heart (electrocardiogram), etc. If two or more variables are involved in the measurement, then the time-series is called a Multivariate Time Series (MTS). For example, in patient monitoring applications, multiple variables such as temperature, pulse rate, blood pressure, and oxygen rate may be involved in monitoring the patient's condition.

Time Series Classification (TSC) is one of the popular research areas over the past few years, mainly due to its numerous practical applications in various domains such as agriculture, healthcare, medicine, finance, and industries [7]. The main objective of TSC is to maximize the prediction accuracy by utilizing the complete sequence data. However, in real scenarios data is collected over time, and thus it is desirable to classify the time-series at an early stage without waiting for full-length sequence data. 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 [8], timely prediction of droughts and shortage of multiple resources would enable the implementation of necessary measures to prevent famine and determine sustainable policy. A classification approach with the aim of classifying the incomplete time series is referred to as early classification [9].

In recent times, early classification of time series has evoked great interest among the researchers [10, 11, 12, 13, 14]. As a result, early classification of time series has shown promising applications such as early disease prediction [15, 16], gas leakage detection [17], electricity demand prediction [18], and drought prediction [8]. The fundamental difference between traditional and early classification approaches has been shown in Figure 1.1 where both models are learned from the same training set. Figure 1.1 - Part(a) indicates the traditional TSC approach in which the classification model classifies the time series when its complete sequence becomes available. Whereas, part (b) demonstrates an early classification approach in which the classification model predicts the class label based on incomplete time series.

A general framework for early classification of time series has been shown in Figure

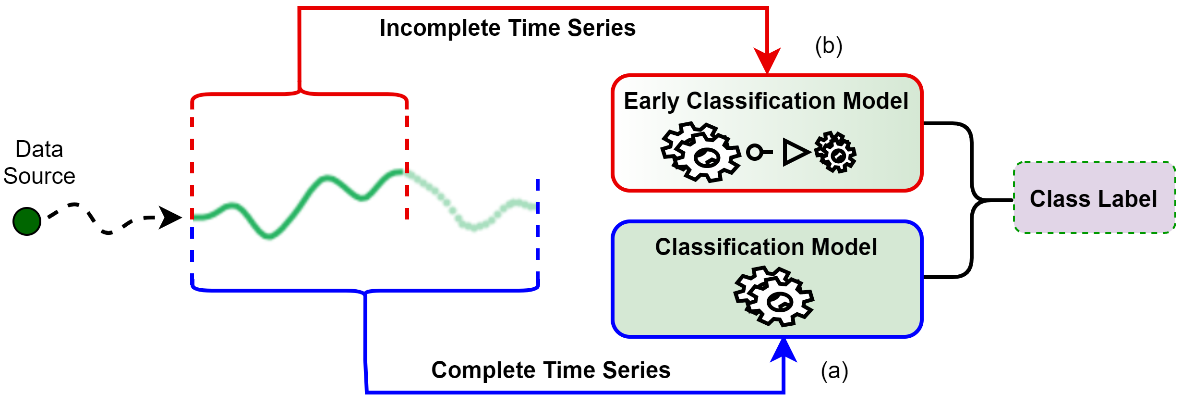

Figure 1.1: Illustration of (a) Traditional TSC and (b) Early classification approach.

1.2. Let's T is the length of complete time series. The early classification framework processes the incoming time series at each time step t and gives it to the classifier. The classifier processes the partially observed data (incomplete time series) and predicts the class label if decision criteria are satisfied. Therefore, the earliness is defined as the number of data points in the incomplete time series at the time of decision. In other words, the minimum number of data points used for early classification. The early classification framework consists of two important components: early classifier and decision criteria. In this regard, the classifier should be adaptable to incomplete time series so that it can produce classification results at different time step t. The decision criteria should be optimized well for making reliable class prediction and classify the time series when adequate information becomes available.

The objective of traditional TSC approach is to classify the time series accurately. Thus it has only one objective that is to maximize accuracy. On the other hand, the aim of an early classification approach is to classify time series as early as possible with desirable accuracy. Henceforth, early classification approach has two objectives, i.e., accuracy and earliness [9]. The general intuition about an early classification problem is that the more data points a time series have, the more reliable perdition it does as it contains more information about the event or activity. Hence, earliness can be

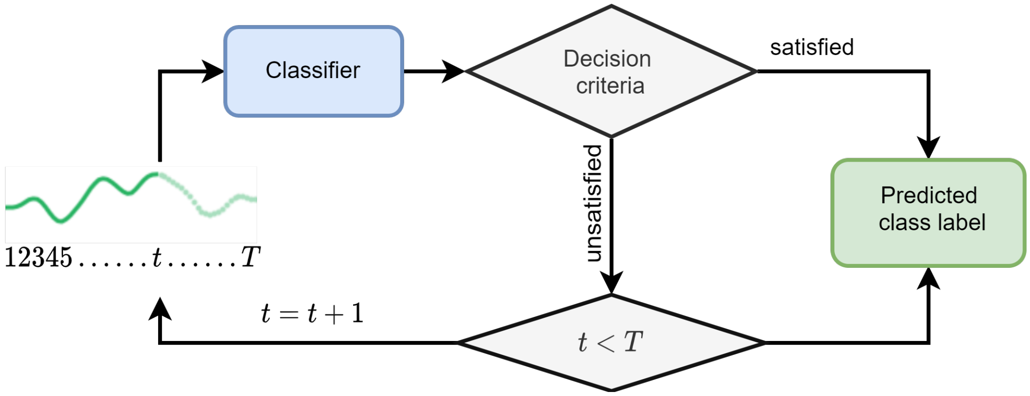

Figure 1.2: A general framework of early classification model

achieved only by compromising with accuracy. So, accuracy and earliness are two conflicting objectives of early classification approaches.

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

The rest of the chapter is organized as follows. The next section presents the motivation of the thesis, followed by the challenges and main objectives. Section 1.3 presents the contribution of this thesis, and Section 1.4 provides the thesis organization details.

1.1 Motivation

Traditional TSC algorithms consider a time series as a whole before predicting its class label [7]. Whereas, in time-sensitive applications, it is highly desirable or even required to predict the class label before the entire series has been observed. For example, in clinical diagnosis, it is often worth sacrificing some classification accuracy in favour of earlier predictions to give clinicians enough time to address infections as they evolve for the sake of the patient's health and to curb the spread of infections. In these settings, an analyst must determine how much accuracy can be sacrificed in favour of earliness with the optimal trade-off depending on the task. Thus early classification approach has great potential to solve time-critical problems in many areas, including security, transportation, and industries, that motivate us to take early classification problems as a challenge.

- **Malware detection in Computer security:** *Malware is one of the major cybersecurity threats in the digital world and is on the rise every day. Basically, it is a malicious program that is deliberately designed to undermine computer security or to harm the computer system like a virus, worm, adware, spyware, trojan, and so on. Behaviour analysis of malware can be detected by running the code in a closely monitored isolated environment or on real systems [19]. Thus, early detection of malicious activity could help in preventing the computer systems from getting harmed from malicious operations.*
- **User mode detection in transportation:** *The transportation modes are the essential part of Intelligent Transportation Systems (ITS) from the user's context that signify how the users are moving around. Transportation mode detection provides valuable support in various ITS applications, such as driving behaviour monitoring [4], [5], human activity monitoring [6], urban transportation planning [7], road environment, and traffic prediction [8] [9], etc. Early detection of transportation modes improves the decision policy in ITS applications.*

Further the motivation of this thesis is derived from the limitations of previously existing works. In early decision making, it is always challenging to identify the optimal time to make a decision because there is always some degree of conflict between accuracy and earliness. Thus, the main challenge in early classification problem is to define the good decision rule by considering both the objectives, i.e., accuracy and earliness, which take part in the early decision process to classify the incoming time series. In literature, an optimal trade-off between accuracy and earliness is defined as a challenge even though this is not considered while designing decision policy by most of the researchers in this domain.

1.2 Challenges and main objectives of this thesis

The development of early classification approach is *challenging* due to following reasons:

- *Building classifier*: Building an early classifier is difficult compared to traditional classification methods. Traditional classification approaches are designed to classify the complete time series. However, an early classifier needs to classify incomplete time series, and therefore it is required to deal with missing values.
- *Decision criteria*: It decides when to stop considering additional information online and make a class prediction, where event by event data points are added in sequence. Here the challenge of decision policy is to pick the right time for predicting a class label without needless delay.
- *Multiple conflicting objectives*: Early classification approaches need to fulfil two objectives: accuracy and earliness, that contradict each other. Maximally inclined early classifier towards the objective (earliness) may not provide accurate prediction due to the non-availability of adequate information. On the other hand, the late classification may cause unnecessary delay in prediction and miss the previous opportunity to react. Thus, balancing between these two objectives is needed.
- *Multivariate signals*: In MTS, class-specific information is developed at different time steps among variables. Moreover, the MTS variables have interconnected relationships, and also these variables may have different lengths. It makes the early classification of MTS more challenging than a single variable signal to design optimal decision policy.

An early classifier should be able to classify the time series at the earliest with reliability so that classification results can be used for further actions. Thus the main *objectives* of the thesis are:

- To build an early classification model that is adaptable to incomplete time series to predict the class label.
- To define optimal decision criteria by considering the trade-off between earliness

and quality in prediction, which can determine whether the partially observed information is sufficient for reliable class prediction or not.

- To achieve the trade-off between the two objectives, i.e., accuracy and earliness.

1.3 Contributions of the thesis

In this thesis, we develop early classification approaches for both UTS as well as MTS data. In our proposed early classification approaches, we have employed generative and discriminative classifiers with optimal decision criteria that have been optimized by balancing the trade-off between accuracy and earliness. In particular, we made the following contributions:

- We propose an early classification approach for UTS by considering the uncertainty in the prediction and also defining optimal decision rules with trade-off optimization between accuracy and earliness.
- Next, we propose the early classification approach for MTS that learns the decision criteria by optimizing the trade-off between accuracy and earliness. The proposed model uses the ensemble classification approach to label the incomplete MTS by capturing the temporal information from each variable separately and then use the collective information for early decision-making. The model learns the decision criteria by optimizing the miss-classification cost and delaying decision cost simultaneously through particle swarm optimization. Moreover, the proposed model employed the Gaussian process probabilistic classifiers to classify the time series.
- Finally, we propose an early classification approach for time series data by developing a hybrid deep learning model. The proposed model uses an imputation-based strategy to classify the incomplete time series and uses a confidence threshold for making the reliable class prediction. The proposed model is able to capture the temporal information of time series without applying any feature engineer-

ing method and defines the confidence threshold as decision criteria by balancing 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 have achieved a decent trade-off between accuracy and earliness. Moreover, the usefulness of the early classification approach has been shown in time-sensitive applications, including malware detection, fault detection, and transportation mode detection.

Some major components of the thesis have been published; particularly, major results in Chapter 3, Chapter 4, and Chapter 5 have been published in references [20], [21] and [22], respectively. A part of chapter 2 has been published as a book chapter in [23].

1.4 Organization of the thesis

This chapter started with the problem definition, followed by motivations, challenges and contributions of the thesis. The rest of this thesis has been organized as follows:

Chapter 2 presents the TSC background and related works of early classification. The background explains the time series representation, similarity measures and classification task with underlying classification strategies. The next part describes an early classification framework with underlying classifier and decision strategies. Finally, related works provide a thorough study of early classification approaches with state-of-the-art methods.

Chapter 3 presents an early classification approach for UTS. The proposed model uses a probabilistic classification approach with two decision policies. In the first part, early decision policies are defined by considering the uncertainty in class prediction. In the second part, decision rules are defined by optimizing accuracy and earliness. We also validate the effectiveness of the proposed approach for early malware detection on publicly available real-world dataset [24, 25].

Chapter 4 extends the optimization-based early classification approach for MTS, presented in Chapter 3. The proposed model learns the decision rules by incorporating the classification results of each variable of MTS, and final class prediction is performed by utilizing the ensemble approach. The results illustrate the significance of the early classification approach using accuracy and earliness.

Chapter 5 introduces an early classification approach adaptable to both UTS as well as MTS. The proposed method develops hybrid deep learning model to design the base classifier and utilized it with imputation-based strategy to handle incomplete time series for classification task. Further decision policy is defined as confidence threshold that facilitates the trade-off between accuracy and earliness.

Chapter 6 summarises the thesis work with promising future research directions in the area of early classification.