Chapter 1

Introduction

With the recent advancements in the embedded systems and communication technologies, sensor-based devices have become a part of our daily routine. These devices generate a temporal sequence of measurements (or data points), which needs to be analyzed for making an appropriate decision. Such a sequence of data points is called as time series [1]. There exists an inherent temporal dependency in the data points which allows the researchers to analyze the behavior of any process over time. In addition to this, the time series has a natural property to satisfy human eagerness of visualizing the structure or shape of the data [2]. Recently, the time series data has received an unprecedented attention in several fields of research, to name a few healthcare [3–5], finance [6, 7], speech and activity recognition [8, 9], and so on [10, 11].

Time Series Classification (TSC) remained a topic of great interest since the availability of labeled dataset repositories such as UCR [12] and UCI [13]. As a consequence, large number of TSC algorithms [1, 14] have emerged by introducing efficient and cutting-edge strategies for distinguishing the classes. The main objective of TSC algorithms is to optimize accuracy of the classification by using complete time series. However, in time-sensitive applications such as gas leakage detection [15], earthquake [16], and electricity demand prediction [17], it is desirable to classify time series

as early as possible without waiting for all data points. A classification approach that aims to classify an incomplete time series is referred as *early classification* [18–20]. Figure 1.1 depicts the difference between the traditional classification and early classification approaches. Both the approaches learn for a training dataset with labeled instances (*i.e.,* time series), but have different capabilities of classification for a testing time series.

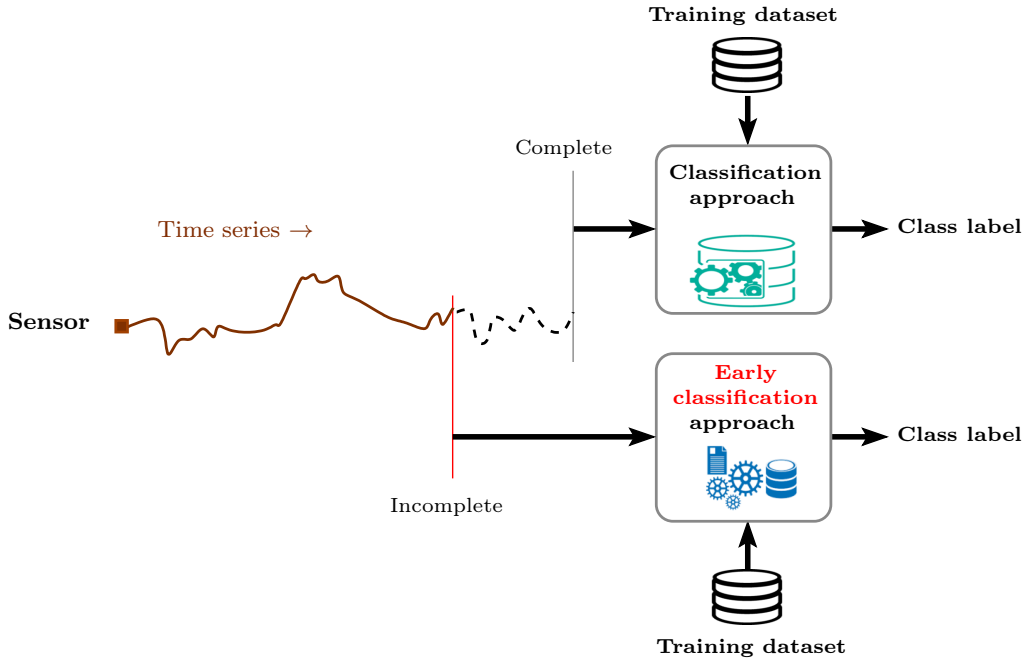

Figure 1.1: Illustration of difference between traditional classification and early classification approaches for time series.

In real-world applications such as human activity recognition and industrial process monitoring, many sensors (*e.g.,* accelerometer, gyroscope, temperature, *etc.*) are used to record subtle information about the activity or process, which helps to obtain highly accurate results. These sensors together generate high dimensional time series data, also called as Multivariate Time Series (MTS) [21, 22]. If a time series is a dimension (or part) of MTS then it can be referred as *component* [23]. In general, a time series is *univariate* unless it is explicitly mentioned as *multivariate*. The early classification of MTS refers to the prediction of class label as early as possible by using minimum

number of data points [8, 24]. The minimum number of data points is acknowledged as Minimum Required Length (MRL) of the MTS. Such MRLs can be learned using MTS training dataset, which help to classify an incomplete MTS at the earliest. Further, the remaining data points that are not used by the classifier, contribute towards earliness. Figure 1.2 illustrates an early classification approach for sensor generated MTS with a given MTS training dataset.

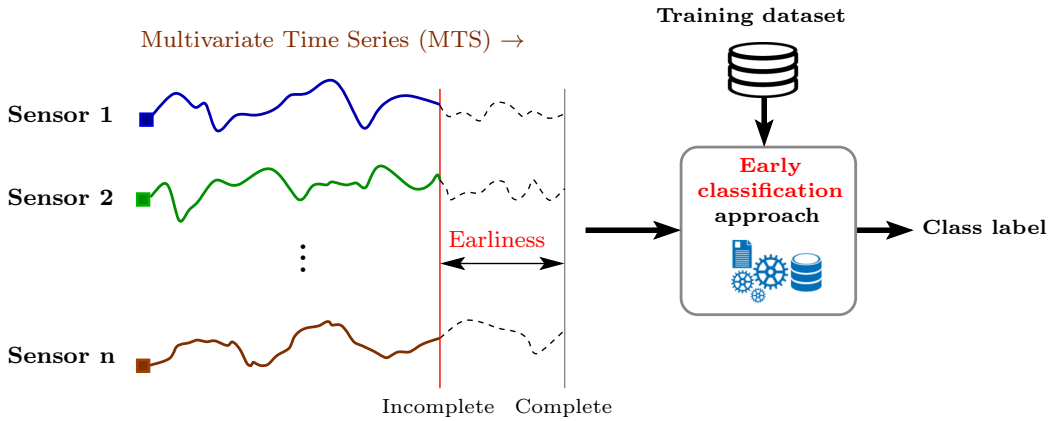

Figure 1.2: Illustration of an early classification approach for MTS.

It is true that the earliness can only be achieved at the cost of accuracy [18, 25]. In other words, if more data points are used in the classification then higher accuracy can be achieved but it will reduce the earliness significantly. Now, the main challenge before an early classification approach is to optimize the tradeoff between two these conflicting objectives, *i.e.,* accuracy and earliness.

This thesis focuses on how to solve early classification problem for sensors generated MTS with the challenges such as different sampling rate of sensors, faulty sensors, and unseen classes. These challenges are application dependent, for example, modern vehicles are embedded with the sensors of different sampling rate to observe inside and outside conditions of the vehicles. The sensors generate an MTS with varying length of components. Another example is human activity classification system where many sensors are used to obtain highly accurate results, however, some of them may

be redundant or faulty. It means the generated MTS may have faulty components. To address the above mentioned challenges, we develop probabilistic approaches for the early classification of MTS while maintaining a desired level of accuracy.

The rest of the chapter is organized as follows: Next section presents the motivation of this thesis. Section 1.2 presents the contribution of the research work and Section 1.3 outlines the organization of the thesis.

1.1 Motivation of the Research Work

In the field of data mining and machine learning, early classification of MTS has received a great attention as it has potential to solve time-critical problems of many areas including healthcare, industry, and transportation. Literature indicates several existing work [5, 22, 23, 26, 27, 27, 28, 28, 29, 29–34] on early classification of MTS that have attempted to improve the earliness with the small compromise of accuracy. The work in this thesis is motivated by the limitations of the existing work as discussed below.

The first limitation is the assumption of the equal length of components in the MTS [22,26–29]. In order to have equal length of components, the sensors must generate the time series at equal sampling rate. Such an assumption is unrealistic in real-world applications where the sensors are purposefully set to have different sampling rate for obtaining fine-grained information without redundancy. An example of such application is intelligent transportation system where different types of sensors are embedded with the vehicles to observe the environmental conditions such as light intensity, temperature, acceleration, vibration, *etc.* Since the embedded sensors are measuring the different conditions, the number of samples taken by the different sensors in a given time period may not be same. As soon as a change occurs in the condition, it should be recorded by the sensor. If the change occurs frequently, then the sampling rate of the sensor should be high enough to capture such frequent changes. On the contrary, if the changes are not frequent then the sampling rate should be low to avoid the collection of redundant

samples. Hence, it is inappropriate to have all the sensors with same sampling rate.

The second limitation of the prior work [5, 23, 30, 31, 35] is the inability to handle faulty data components of the MTS. The faulty data components may be present due to faulty or unreliable sensors. In real-world applications such as human activity classification, multiple sensors are used to obtain correct information about the activity being performed. However, the correctness of information depends on the reliability of sensors, which can be assessed using the generated data [36]. Further, highly accurate results can be obtained by increasing the number of sensors but it will also increase the computational complexity of the classifier. In a resource constrained environment (*e.g.,* smartphone and wearable), it is desirable to optimize the computational complexity while maintaining a desired level of accuracy of classification. It indicates that the time series corresponding to some sensors can be omitted if they contribute marginally in the accuracy [37]. Such time series (components) of MTS can be removed by considering them as faulty during early classification. It indicates the demand of a fault-tolerant early classification approach for MTS.

Another major limitation of the existing approaches [27–29, 32–34] is that they can classify an incomplete MTS correctly only if it belongs to a known or seen class label. A class label is said to be seen if it appears in the training dataset otherwise it is called as unseen [38]. However, in some applications such as appliance monitoring, early classification of the MTS of unseen class is desirable for fault identification. For example, with the advent of sensors in industrial and domestic applications, it becomes easy to diagnose the abnormal or faulty behavior of the appliances using sensory data. Faulty operation of an appliance causes significant fluctuations in the sensory measurements. Such fluctuations help to distinguish a faulty operation from the normal [39, 40]. An appliance may encounter several types of faults during its lifetime. However, it is not practical to have prior knowledge (*i.e.,* MTS in the training dataset) about all types of fault. It indicates that the existing approaches are not suitable for fault classification in

the industrial or domestic appliances as an unseen fault may occur at any time during its operation.

In addition to above limitations, some prior work [23, 26, 32, 41] do not consider the correlation among the components of MTS while predicting its class label. The correlation can be quiet informative for identifying the class label at early stage of the MTS and thus it can help to achieve better earliness. The correlation is inherently present among the components of MTS if the sensors are observing the same phenomenon. For example, in autonomous vehicles, the driving pattern of a driver can be identified by using the on-vehicle accelerometer, gyroscope, and pressure sensors, where accelerometer and gyroscope capture the motion related information and pressure sensor observes how much force is applied by the driver on the pedals (*i.e.,* accelerator, break, and clutch). These sensors together generate an MTS of the driving pattern where correlation should necessarily be present. Avoiding such correlation hampers the accuracy of classification which could be achieved at early stage of the MTS.

1.2 Contributions of the Thesis

In this thesis, we develop early classification approaches to classify a sensors generated MTS while maintaining a desired level of accuracy. Essentially, we address the challenges of MTS such as different length of components, faulty data components, and unseen class labels. We employ probabilistic classifier (*e.g.,* Gaussian Process [42]) to build early classifiers by learning class-wise MRLs with tradeoff optimization. With a given training MTS dataset, this thesis aims to answer the following research questions:

- How to classify an incomplete MTS with different length of components that are generated from the sensors of different sampling rate?
- How early the class label of an MTS can be predicted when some of its components are generated from faulty or unreliable sensors?
- How to classify an incomplete MTS if it belongs to an unseen class label for which

there exists no instance in the training dataset?

- As correlation among the components of MTS can be quite informative for predicting the class label of an incomplete MTS, how to incorporate such correlation in the early classification approach?

1.3 Organization of the Thesis

The rest of the thesis is organized as follows.

Chapter 2: This chapter presents a systematic review of current literature on early classification approaches for MTS. We review and categorize the existing approaches based on the strategies that they have followed for obtaining the earliness.

Chapter 3: In this chapter, we propose an early classification approach for MTS with a desired level of accuracy. *It is assumed that the number of samples in the components are not equal for a given period of time due to different sampling rate of sensors.* Gaussian Process (GP) classifier is used to first estimate the minimum required length of the time series which helps to build an ensemble classifier with a desired level of accuracy. The ensemble classifier is used to predict the class label of an incomplete MTS. This chapter also demonstrates a road surface classification system using the built ensemble classifier. Finally, the ensemble classifier is evaluated on the various existing datasets from other domains. The results illustrate the significance of the early classification approach using accuracy, earliness, and confusion matrix, with the minimum required data points.

Chapter 4: In this chapter, we propose a Fault-tolerant Early Classification of MTS (FECM) approach *to address the presence of faulty data components in the MTS.* FECM

builds a set of classification models using MTS training dataset. The approach employs GP classifier to estimate minimum required length of components, which is used to predict a class label of a new incomplete MTS. Later, FECM uses an Auto Regressive Integrated Moving Average (ARIMA) model to identify faulty components in the incomplete MTS. Finally, we evaluate the performance of FECM for early classification of the human activities such as standing while talking, sitting on sofa, eating, and so on.

Chapter 5: In this chapter, we propose a semantic-information based early classification approach for MTS. The proposed approach uses a concept of Zero-Shot Learning *to classify an MTS of unseen class by utilizing the most identifiable attributes (semantic information) of the seen classes.* This chapter also conducts a case study to evaluate the proposed approach by classifying various types of faults of the washing machine using sensory data. In addition, we validate the effectiveness of the approach on two real-world existing datasets obtained from UCI repository [13].

Chapter 6: This chapter summarizes the thesis with main findings of the research work. We also discuss some promising directions for conducting future research in the area of early classification.

Some important and relevant research papers are listed in References. At the end, we provide the List of Publications from the research work presented in this thesis.