IGI PUBLISHING **ITB15224**

701 E. Chocolate Avenue, Suite 200, Hershey PA 17033-1240, USA
Tel: 717/533-8845; Fax 717/533-8661; URL-http://www.igi-global.com

This paper appears in the publication, **Data Mining and Knowledge Discovery Technologies** edited by D. Taniar © 2008, IGI Global

Chapter VII

Advances in Classification of Sequence Data

Pradeep Kumar, University of Hyderabad, Gachibowli, India*
P. Radha Krishna, University of Hyderabad, Gachibowli, India
Raju S. Bapi, University of Hyderabad, Gachibowli, India
T. M. Padmaja, University of Hyderabad, Gachibowli, India

Abstract

In recent years, advanced information systems have enabled a collection of increasingly large amounts of data that are sequential in nature. To analyze huge amounts of sequential data, the interdisciplinary field of knowledge discovery in databases (KDD) is very useful. The most important step within the process of KDD is data mining, which is concerned with the extraction of the valid patterns. Recent research focus in data mining includes stream data mining, sequence data mining, Web mining, text mining, visual mining, multimedia mining, and multi-relational data mining. Sequence data may be discrete or continuous in nature. Most of the research on discrete sequence data concentrated on the discovery of frequently occurring patterns. However, comparatively less amount of work has been carried out in the area of discrete sequence data classification. In this chapter, data taxonomy is introduced with a review of the state of art for sequence data classification. The usefulness of embedding partial subsequence information extracted using sliding window technique into traditional classifier like kNN has been demonstrated. kNN

has been tested with various vector based distance/similarity metrics. Further, with the use of S^3M similarity metric, the full subsequence information embedded in the data sequences is extracted. The experimental data taken is DARPA'98 IDS benchmark dataset collected from UCIML dataset repository. The chapter closes by pointing out various application areas of sequence data and also the open issues in sequence data classification problem.

Introduction

In recent years, the amount of data that is collected by advanced information systems has increased tremendously. The resulting data volume is too large to be examined manually and even the methods for automatic data analysis based on classical statistics and machine learning often face problems when processing large, dynamic data sets consisting of complex data. The interdisciplinary field of knowledge discovery in databases (KDD) helps in analyzing these large volumes of data. KDD employs methods at the cross-post of machine learning, statistics, and database systems. Fayyad, Shapiro, and Smyth (1996) define KDD as follows:

Knowledge discovery in databases is the non-trivial process of identifying valid, novel, potentially useful, and ultimately understandable patterns in data.

According to this definition, data is a set of facts that is somehow accessible in electronic form. The term patterns indicate models and regularities, which can be observed within the data. Patterns have to be valid (i.e., they should be true on new data with some degree of certainty). A novel pattern is not previously known or trivially true. The potential usefulness of patterns refers to the possibility that they lead to an action providing a benefit. A pattern is understandable if it is interpretable by a human user. KDD is a process comprising several steps that are perhaps repeated over several iterations.

The core step of KDD is called *data mining*. Data mining can be defined as an activity that extracts new and nontrivial information contained in large databases. The goal is to discover hidden patterns, unexpected trends or other non-obvious relationships in the data using a combination of techniques such as, machine learning, statistics, and databases.

A huge amount of data is collected every day in the form of sequences. These sequential data are valuable sources of information not only to search for a particular value or event at a specific time, but also to analyze the frequency of certain events or sets of events related by particular temporal/sequential relationship. Examples