Chapter 38
Quantization based Sequence Generation and Subsequence Pruning for Data Mining Applications

T. Ravindra Babu
Infosys Limited, India

M. Narasimha Murty
Indian Institute of Science Bangalore, India

S. V. Subrahmanya
Infosys Limited, India

ABSTRACT

Data Mining deals with efficient algorithms for dealing with large data. When such algorithms are combined with data compaction, they would lead to superior performance. Approaches to deal with large data include working with representatives of data instead of entire data. The representatives should preferably be generated with minimal data scans. In the current chapter we discuss working with methods of lossy and non-lossy data compression methods combined with clustering and classification of large datasets. We demonstrate the working of such schemes on two large data sets.

INTRODUCTION

With increasing number of transactions, reducing cost of storage devices, and the need for generating abstractions for business intelligence, it has become important to search for efficient methods for dealing with large, sequential and time series data. Data mining (Agrawal, et al, 1993; Fayyad, et al, 1996; Han & Kamber, 1996) focuses on development of scalable and efficient generation of valid, general and novel abstraction from a large dataset.

A transactional dataset consists of records that have transaction-id and the items that make up the transaction. A temporal dataset stores relational data that included time-related attri-

DOI: 10.4018/978-1-4666-2455-9.ch038

butes. A sequence dataset contains sequences of ordered events, with or without time information. A time-series dataset contains sequences of values or events obtained over repeat measurements of time periodically like those of spacecraft health data, data from stock exchange, etc. Data Mining is inter-disciplinary subject that encompasses a number of disciplines like Machine Learning, large data clustering and classification, statistics, algorithms, etc.

In the current chapter, we present schemes for non-lossy and lossy compression of data using sequence generation, run-length computation, subsequence pruning leading to efficient clustering and classification of large data. The schemes are efficient, scale up well and provide high classification accuracy.

The proposed scheme integrates the following.

A. Vector Quantization
B. Sequence Generation
C. Item Support and Frequent subsequences (Agrawal et al., 1993; Han et al., 2000)
D. Subsequence Pruning (Ravindra, Murty, & Agrawal, 2004)
E. Run length encoding
F. Support Vector Machines
G. Classification

The chapter is organized into sections. We discuss motivation for the work in the following section. It is followed by discussion on related work, background terminology and concepts along with illustrations. It is followed by a description of datasets on which we deomstrate working of the proposed schemes. The description includes summary of preliminary analysis of the datasets. Then the following section contains a discussion on proposed scheme, experimentation and results followed by a section on discussion on future research directions. Finally the work is summarized in the last section.

Motivation

When data is large, operating on every pattern to generate an abstraction is expensive both in terms of space and time. In addition, as the data size increases, multiple scans of database would become prohibitive. Hence, generation of abstraction should happen in a small number of scans, ideally a single scan.

Some approaches to deal with large and high dimensional data make use of optimal representative patterns or optimal feature set to represent each pattern. Alternatively, it is interesting to explore whether it is possible to deal with data by compressing the data and work in the compressed domain without having to decompress.

Compression would lead to reduction in space requirements. Further it is also interesting to explore, while compressing the data, whether we can work only on subset of features based on some criterion. This would lead to working in lossy compression domain. However care should be exercised in ensuring that the necessary information is not lost in the process.

We propose two such schemes and examine whether such schemes work efficiently on large datasets in terms of pattern classification.

BACKGROUND

Related Literature

Large data clustering schemes (Jain, Murty & Flynn, 1999) provide ways to deal with large data. Some of the successful methods in this direction have been optimal prototype selection schemes (Ravindra & Murty, 2001; Susheela, 2010), multi-agent systems for large data clustering (Ravindra, Murty & Subrahmanya, 2010; Ravindra, Murty & Subrahmanya, 2009), optimal feature selection (Kim, Street & Mericzer, 2003) and simultaneous selection of prototype and features (Ravindra, Murty & Agrawal, 2005).