Instance Selection

Huan Liu
Arizona State University, USA

Lei Yu
Arizona State University, USA

INTRODUCTION

The amounts of data have become increasingly large in recent years as the capacity of digital data storage worldwide has significantly increased. As the size of data grows, the demand for data reduction increases for effective data mining. Instance selection is one of the effective means to data reduction. This article introduces the basic concepts of instance selection and its context, necessity, and functionality. The article briefly reviews the state-of-the-art methods for instance selection.

Selection is a necessity in the world surrounding us. It stems from the sheer fact of limited resources, and data mining is no exception. Many factors give rise to data selection: Data is not purely collected for data mining or for one particular application; there are missing data, redundant data, and errors during collection and storage; and data can be too overwhelming to handle. Instance selection is one effective approach to data selection. This process entails choosing a subset of data to achieve the original purpose of a data-mining application. The ideal outcome is a model independent, minimum sample of data that can accomplish tasks with little or no performance deterioration.

BACKGROUND AND MOTIVATION

When we are able to gather as much data as we wish, a natural question is “How do we efficiently use it to our advantage?” Raw data is rarely of direct use, and manual analysis simply cannot keep pace with the fast accumulation of massive data. Knowledge discovery and data mining (KDD), an emerging field comprising disciplines such as databases, statistics, and machine learning, comes to the rescue. KDD aims to turn raw data into nuggets and create special edges in this ever-competitive world for science discovery and business intelligence. The KDD process is defined as the nontrivial process of identifying valid, novel, potentially useful, and ultimately understandable patterns in data (Fayyad, Piatetsky-Shapiro, Smyth, & Uthurusamy, 1996). It includes data selection, preprocessing, data mining, interpretation, and evaluation. The first two processes (data selection and preprocessing) play a pivotal role in successful data mining (Han & Kamber, 2001). Facing the mounting challenges of enormous amounts of data, much of the current research concerns itself with scaling up data-mining algorithms (Provost & Kolluri, 1999). Researchers have also worked on scaling down the data — an alternative to the scaling up of the algorithms. The major issue of scaling down data is to select the relevant data and then present it to a data-mining algorithm. This line of work is parallel with the work on scaling up algorithms, and the combination of the two is a two-edged sword in mining nuggets from massive data.

In data mining, data is stored in a flat file and described by terms called attributes or features. Each line in the file consists of attribute-values and forms an instance, which is also called a record, tuple, or data point in a multidimensional space defined by the attributes. Data reduction can be achieved in many ways (Liu & Motoda, 1998; Blum & Langley, 1997; Liu & Motoda, 2001). By selecting features, we reduce the number of columns in a data set; by discretizing feature values, we reduce the number of possible values of features; and by selecting instances, we reduce the number of rows in a data set. We focus on instance selection here.

Instance selection reduces data and enables a data-mining algorithm to function and work effectively with huge data. The data can include almost everything related to a domain (recall that data is not solely collected for data mining), but one application normally involves using one aspect of the domain. It is natural and sensible to focus on the relevant part of the data for the application so that the search is more focused and mining is more efficient. Cleaning data before mining is often required. By selecting relevant instances, we can usually remove irrelevant, noise, and redundant data. The high-quality data will lead to high-quality results and reduced costs for data mining.

Copyright © 2006, Idea Group Inc., distributing in print or electronic forms without written permission of IGI is prohibited.
Related Content

Physical Modeling of Data Warehouses Using UML Component and Deployment Diagrams: Design and Implementation Issues  
www.igi-global.com/chapter/physical-modeling-data-warehouses-using/7665?camid=4v1a

Knowledge Discovery in Databases with Diversity of Data Types  
www.igi-global.com/chapter/knowledge-discovery-databases-diversity-data/10961?camid=4v1a

Ensemble Data Mining Methods  
www.igi-global.com/chapter/ensemble-data-mining-methods/10639?camid=4v1a

Analysis of Content Popularity in Social Bookmarking Systems  
www.igi-global.com/chapter/analysis-content-popularity-social-bookmarking/38226?camid=4v1a