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.

MAJOR LINES OF RESEARCH AND DEVELOPMENT

A spontaneous response to the challenge of instance selection is, without fail, some form of sampling. Although sampling is an important part of instance selection, other approaches do not rely on sampling but resort to search or take advantage of data-mining algorithms. In this section, we start with sampling methods and proceed to other instance-selection methods associated with data-mining tasks, such as classification and clustering.

Sampling Methods

Sampling methods are useful tools for instance selection (Gu, Hu, & Liu, 2001).

Simple random sampling is a method of selecting n instances out of the N such that every one of the $\binom{N}{n}$ distinct samples has an equal chance of being drawn. If an instance that has been drawn is removed from the data set for all subsequent draws, the method is called *random sampling without replacement.* Random sampling *with* replacement is entirely feasible: At any draw, all N instances of the data set have an equal chance of being drawn, no matter how often they have already been drawn.

Stratified random sampling divides the data set of N instances into subsets of $N_1, N_2, \ldots, N_l$ instances, respectively. These subsets are nonoverlapping, and together they comprise the whole data set (i.e., $N_1 + N_2, \ldots, +N_l = N$). The subsets are called *strata*. When the strata have been determined, a sample is drawn from each stratum, the drawings being made independently in different strata. If a simple random sample is taken in each stratum, the whole procedure is described as stratified random sampling. It is often used in applications when one wishes to divide a heterogeneous data set into subsets, each of which is internally homogeneous.

Adaptive sampling refers to a sampling procedure that selects instances depending on results obtained from the sample. The primary purpose of adaptive sampling is to take advantage of data characteristics in order to obtain more precise estimates. It takes advantage of the result of preliminary mining for more effective sampling, and vice versa.

Selective sampling is another way of exploiting data characteristics to obtain more precise estimates in sampling. All instances are first divided into partitions according to some homogeneity criterion, and then random sampling is performed to select instances from each partition. Because instances in each partition are more similar to each other than instances in other partitions, the resulting sample is more representative than a randomly generated one. Recent methods can be found in Liu, Motoda, and Yu (2002), in which samples selected from partitions based on data variance result in better performance than samples selected with random sampling.

Methods for Labeled Data

One key data-mining application is *classification* — predicting the class of an unseen instance. The data for this type of application is usually labeled with class values. Instance selection in the context of classification has been attempted by researchers according to the classifiers being built. In this section, we include five types of selected instances.

Critical points are the points that matter the most to a classifier. The issue originated from the learning method of Nearest Neighbor (NN) (Cover & Thomas, 1991). NN usually does not learn during the training phase. Only when it is required to classify a new sample does NN search the data to find the nearest neighbor for the new sample, using the class label of the nearest neighbor to predict the class label of the new sample. During this phase, NN can be very slow if the data are large and can be extremely sensitive to noise. Therefore, many suggestions have been made to keep only the critical points, so that noisy ones are removed and the data set is reduced. Examples can be found in Yu, Xu, Ester, and Kriegel (2001) and Zeng, Xing, and Zhou (2003), in which critical data points are selected to improve the performance of collaborative filtering.

Boundary points are the instances that lie on borders between classes. Support vector machines (SVM) provide a principled way of finding these points through minimizing structural risk (Burges, 1998). Using a nonlinear function $\varnothing$ to map data points to a high-dimensional feature space, a nonlinearly separable data set becomes linearly separable. Data points on the boundaries, which maximize the margin band, are the support vectors. Support vectors are instances in the original data sets and contain all the information a given classifier needs for constructing the decision function. Boundary points and critical points are different in the ways they are found.

Prototypes are representatives of groups of instances via averaging (Chang, 1974). A prototype that represents the typicality of a class is used in characterizing a class rather than describing the differences between classes. Therefore, they are different from critical points or boundary points.

Tree-based sampling is a method involving decision trees (Quinlan, 1993), which are commonly used classification tools in data mining and machine learn-

Related Content

Development of Control Signatures with a Hybrid Data Mining and Genetic Algorithm

Alex Burns, Shital Shahand Andrew Kusiak (2008). *Data Warehousing and Mining: Concepts, Methodologies, Tools, and Applications (pp. 2226-2247).*

www.irma-international.org/chapter/development-control-signatures-hybrid-data/7757

Continuous Auditing and Data Mining

Edward J. Garrity, Joseph B. O'Donnelland G. Lawrence Sanders (2005). *Encyclopedia of Data Warehousing and Mining (pp. 217-222).*

www.irma-international.org/chapter/continuous-auditing-data-mining/10596

Mining in Spatio-Temporal Databases

Junmei Wang, Wynne Hsuand Mong Li Lee (2008). *Data Warehousing and Mining: Concepts, Methodologies, Tools, and Applications (pp. 3477-3492).*

www.irma-international.org/chapter/mining-spatio-temporal-databases/7844

Video Data Mining

Jung Hwan Oh, Jeong Kyu Leeand Sae Hwang (2008). *Data Warehousing and Mining: Concepts, Methodologies, Tools, and Applications (pp. 1631-1637).*

www.irma-international.org/chapter/video-data-mining/7720

Administering and Managing a Data Warehouse

James E. Yao, Chang Liu, Qiyang Chenand June Lu (2008). *Data Warehousing and Mining: Concepts, Methodologies, Tools, and Applications (pp. 1-8).*

www.irma-international.org/chapter/administering-managing-data-warehouse/7627