Chapter II

A Disk-Based Algorithm for Fast Outlier Detection in Large Datasets

Faxin Zhao, Northeastern University, and Tonghua Teachers College, China

Yubin Bao, Northeastern University, China

Huanliang Sun, Northeastern University, China

Ge Yu, Northeastern University, China

Abstract

In data mining fields, outlier detection is an important research issue. The number of cells in the cell-based disk algorithm increases exponentially. The performance of this algorithm will decrease dramatically with the increasing of the number of cells and data points. Through further analysis, we find that there are many empty cells that are useless to outlier detection. So this chapter proposes a novel index structure, called CD-Tree, in which only non-empty cells are stored, and a cluster technique is adopted to store the data objects in the same cell into linked disk pages. Some experiments are made to test the performance of the proposed algorithms.
The experimental results show that the performance of the CD-Tree structure and of the cluster technique based disk algorithm outperforms that of the cell-based disk algorithm, and the dimensionality processed by the proposed algorithm is higher than that of the old one.

Introduction

Outlier detection is an important research issue in data mining. Compared to association rules (Agrawal, Imielinski, & Swami, 1993), classification (Rastogi & Shim, 1998) and clustering (Ng & Han, 1994; Zhang, Ramakrishnan, & Livny, 1996), it aims to discover “Small Pattern” or outliers — the data objects that are dissimilar or inconsistent with the remainder of the data. Hawkins’ definition captures the spirit of outlier: “an outlier is an observation that deviates so much from other observations as to arouse suspicious that it was generated by a different mechanism” (Hawkins, 1980). In many applications, exceptional patterns are more significant than general patterns, such as intrusion detection, telecom and credit card fraud detection, loan approval, and weather prediction.

Some researchers have designed many algorithms for outlier detection based on various assumptions of outliers existing. Knorr and Ng (1998, 1999) proposed a cell-based algorithm for outliers detection, which quantizes each of data objects into a $k$-D ($k$ is equal to the dimensionality of objects in the dataset) space that has been partitioned into cells. Outliers detection is based on these cells, the algorithm has a complexity of $O(c^k + N)$ which is linear wrt $N$, but exponential wrt $k$, $k$ being the dimensionality and $N$ being the number of objects in the dataset. The algorithm doesn’t work effectively for high dimensionality or fine granularity partition because the number of cells is too many. We find that all cells generated by the algorithm are stored and processed. In fact, there exist plenty of empty cells in the partitions of a dataset, and these empty cells are useless for outlier detection except wasting memory and processing time. So, in this chapter, in order to solve the above problems, we present an efficient index structure named CD-Tree (Cell Dimension Tree), which only stores non-empty cells. Based on the index structure CD-Tree, we propose a disk-based algorithm for fast outlier detection in large datasets, and then a cluster technique is also employed to improve the efficiency of the algorithm further. The experimental results show that the performance of the CD-Tree and cluster technique based disk algorithm outperforms that of the cell-based disk algorithm, and the number of dimensions processed by the proposed algorithms is higher than that of the old one.

The remainder of this chapter is organized as follows: Related work work is discussed. The cell-based disk algorithm for outlier detection (CS-d) is introduced. Structure and
Related Content

Multimedia Databases
www.irma-international.org/chapter/multimedia-databases/11178/

Implicit Semantics Based Metadata Extraction and Matching of Scholarly Documents
www.irma-international.org/article/implicit-semantics-based-metadata-extraction-and-matching-of-scholarly-documents/211912/

Alliance Project: Digital Kinship Database and Genealogy
www.irma-international.org/chapter/alliance-project-digital-kinship-database/7952/

Control-Based Database Tuning Under Dynamic Workloads
www.irma-international.org/chapter/control-based-database-tuning-under/8051/

Lock-Based Concurrency Control in Distributed Real-Time Database Systems
Özgür Ulusoy† (1993). Journal of Database Management (pp. 3-17).
www.irma-international.org/article/lock-based-concurrency-control-distributed/51117/