Chapter 8
Online Clustering and Outlier Detection

Baoying Wang
Waynesburg University, USA

Aijuan Dong
Hood College, USA

ABSTRACT

Clustering and outlier detection are important data mining areas. Online clustering and outlier detection generally work with continuous data streams generated at a rapid rate and have many practical applications, such as network instruction detection and online fraud detection. This chapter first reviews related background of online clustering and outlier detection. Then, an incremental clustering and outlier detection method for market-basket data is proposed and presented in details. This proposed method consists of two phases: weighted affinity measure clustering (WC clustering) and outlier detection. Specifically, given a data set, the WC clustering phase analyzes the data set and groups data items into clusters. Then, outlier detection phase examines each newly arrived transaction against the item clusters formed in WC clustering phase, and determines whether the new transaction is an outlier. Periodically, the newly collected transactions are analyzed using WC clustering to produce an updated set of clusters, against which transactions arrived afterwards are examined. The process is carried out continuously and incrementally. Finally, the future research trends on online data mining are explored at the end of the chapter.

1. INTRODUCTION

With the widespread use of network, online clustering and outlier detection as the main data mining tools have drew attention from many practical applications, especially in areas where detecting abnormal behaviors is critical, such as online fraud detection, network instruction detection, and customer behavior analysis. These applications often generate a huge amount of data at a rather rapid rate. Manual screening or checking of this massive data collection is time consuming and impractical. Because of this, online clustering and outlier detection is a promising approach for

DOI: 10.4018/978-1-4666-2455-9.ch008

such applications. Specifically, data mining tools are used to group online activities or transactions into clusters and to detect the most suspicious entries. The clusters are used for marketing and management analysis. The most suspicious ones are investigated further to determine whether they are truly outlier.

Numerous clustering and outlier detection algorithms have been developed (Agyemang, Barker, & Alhajj, 2006; Weston, Hand, Adams, Whitrow, & Juszczak, 2008; Dorronsoro, Ginel, Sgnchez, & Cruz, 1997; Bolton & Hand, 2002; Panigrahi, 2009; He, Deng, & Xu, 2005; Wei, Qian, Zhou, Jin, & Yu, 2003; Aggarwal, Han, Wang, & Yu, 2006; Elahi, Li, Nisar, Lv, & Wang, 2008), but the majority of them are intended for continuous data. With the few approaches for categorical data (He et al., 2005; Wei et al., 2003), time efficiency and detection accuracy need to be further improved. In this chapter, we present an efficient dynamic clustering and outlier detection method for online market basket data. Market basket data are usually organized horizontally in the form of transactions, with each transaction containing a list of items bought (and/or a list of behaviors performed) by a customer during a single checkout at a (online) store. Unlike traditional data, market-basket data are known to be high dimensional, sparse, and to contain attributes of categorical nature.

Our incremental clustering and outlier detection approach consists of two phases: weighted affinity measure clustering (WC clustering) and outlier detection. First, the transaction sets are analyzed so that items are grouped using WC clustering. Then, each newly arrived transaction is examined against the item clusters that are formed in the WC clustering phase. Phase two decides whether the new transaction is an outlier. After a period of time, the newly collected transactions or data streams are analyzed using WC clustering to produce an updated item clusters, against which each newly arrived transaction afterwards is examined. The process continues incrementally.

This proposed online clustering and outliner detection method has the following characteristics:

1. It is incremental. Each newly arrived transaction is examined immediately against the results from the past transactions.
2. The results of WC clustering are item clusters rather than transaction clusters so that the newly arrived transaction is examined against the item clusters rather than the whole past transactions. The number of item clusters is usually much smaller than the number of past transactions clusters.
3. The item clusters are updated periodically so that any new items and any new purchase behaviors of customers are taken into consideration to produce more accurate results for the future detection.
4. Finally, WC affinity measure, developed in our previous work, is used to improve the clustering results hence outlier detection results.

The rest of the chapter is organized as follows. Section 2 introduces background information and reviews previous research in related areas. Section 3 presents the proposed online clustering and outlier detection method in details. Section 4 concludes the research and highlights the future research trend.

2. BACKGROUND

Since the proposed method is an online clustering and outlier detection method that uses vertical data structure and weighted confidence affinity measure, we present a brief literature overview on the following aspects in this section: clustering methods, outlier detection, online data mining, affinity measure between clusters, and vertical data structures.