

Chapter LIX

Outlying Subspace Detection for High-Dimensional Data

Ji Zhang

CSIRO Tasmanian ICT Centre, Australia

Qigang Gao

Dalhousie University, Canada

Hai Wang

Saint Mary's University, Canada

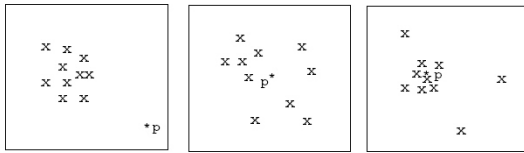
INTRODUCTION

Knowledge discovery in databases, commonly referred to as data mining, has attracted enormous research efforts from different domains such as databases, statistics, artificial intelligence, data visualization, and so forth in the past decade. Most of the research work in data mining such as clustering, association rules mining, and classification focus on discovering large patterns from databases (Ramaswamy, Rastogi, & Shim, 2000). Yet, it is also important to explore the small patterns in databases that carry valuable information about the interesting abnormalities. Outlier detection is a research problem in small-pattern mining in databases. It aims at finding a specific number of objects that are considerably dissimilar, exceptional, and inconsistent with respect to the majority records in an input database. Numerous research

work in outlier detection has been proposed such as the distribution-based methods (Barnett & Lewis, 1994; Hawkins, 1980), the distance-based methods (Angiulli & Pizzuti, 2002; Knorr & Ng, 1998, 1999; Ramaswamy et al.; Wang, Zhang, & Wang, 2005), the density-based methods (Breuning, Kriegel, Ng, & Sander, 2000; Jin, Tung, & Han, 2001; Tang, Chen, Fu, & Cheung, 2002), and the clustering-based methods (Agrawal, Gehrke, Gunopulos, & Raghavan, 1998; Ester, Kriegel, Sander, & Xu, 1996; Hinneburg & Keim, 1998; Ng & Han, 1994; Sheikholeslami, Chatterjee, & Zhang, 1999; J. Zhang, Hsu, & Lee, 2005; T. Zhang, Ramakrishnan, & Livny, 1996).

One important characteristic of outliers in high-dimensional data sets is that they are usually embedded in lower dimensional feature subspaces, and different data points may be considered as outliers in rather different subspaces. To better demonstrate the motivation of exploring outlying

Figure 1. Two-dimensional views of a high-dimensional data space



subspaces, let us consider the example in Figure 1, in which three two-dimensional views of a high-dimensional data space are presented. Note that point p exhibits different outlier qualities in these three views. In the leftmost view, p is clearly an outlier. However, in the middle view, p has a much weaker outlier status and is not an outlier at all in the rightmost view.

The conventional methods of outlier mining, as mentioned above, are mainly designed to detect a certain number of top outliers in a prespecified feature subspace. Consequently, this may render them to miss many outliers hidden in other feature subspaces. It would be computationally prohibitive for them to perform outlier mining in each possible subspace of a high-dimensional feature space. Thus, identifying the subspaces in which each data point is considered as an outlier would be crucial to outlier detection in high-dimensional databases.

This entry focuses on the problem of outlying subspace detection for high-dimensional data. This challenging problem has recently been identified as a subdomain of outlier mining in databases (J. Zhang, Lou, Ling, & Wang, 2004; J. Zhang & Wang, 2006; Zhu, Kitagawa & Faloutsos, 2005). Outlier mining can benefit from outlying subspace detection in the following aspects.

- Outlying subspace detection can enable outlier mining to be performed more accurately. It can allow outliers to be detected from more than one subspace. This is important to many applications. There are abundant examples of high-dimensional data such as spatial

data and gene expression (microarray) data. Outlier detection from these data sets can discover potentially useful abnormal patterns. As we have mentioned earlier, outliers in these high-dimensional data sets are usually hidden in some low-dimensional subspaces. The conventional outlier detection methods may not be able to provide satisfactory effectiveness to these applications. This is because they usually rely on a prespecified feature subspace to detect outliers, which may cause them to miss many potential outliers hidden in other feature subspaces and fail to raise necessary alarms. Therefore, outlying subspace analysis plays a crucial role for outlier mining in these applications to identify the correct feature subspaces in which outliers can be accurately mined. An object is considered as an outlier in these applications if it exhibits abnormality in one or more subspaces.

- Outlying subspace detection can help outlier detection methods to mine outliers in high-dimensional spaces more efficiently. Outlying subspace detection is an important intermediate step in example-based outlier mining in a high-dimensional feature space. Zhu et al. (2005) and Zhu, Kitagawa, Papadimitriou, and Faloutsos (2004) proposed a method to detect outliers based on a set of outlier examples. The basic idea of this method is to find the outlying subspaces of these outlier examples, from which more outliers that have similar outlier characteristics to the given examples can thereby be found in a more efficient manner. This is because outlier detection only needs to be performed in those detected outlying subspaces and the computation cost in other irrelevant subspaces can thus be saved, which leads to a substantial performance gain.
- Outlying subspace detection can contribute to a better characterization of the outliers detected. The characterization of outliers mainly involves presenting the subspaces in which these outliers exist. In a high-dimen-

5 more pages are available in the full version of this document, which may be purchased using the "Add to Cart" button on the publisher's webpage:

www.igi-global.com/chapter/outlying-subspace-detection-high-dimensional/20740

Related Content

The Effects of an Enterprise Resource Planning (ERP) Implementation on Job Characteristics: A Study Using the Hackman and Oldham Job Characteristics Model

Gerald Grant and Aareni Uruthirapathy (2003). *ERP & Data Warehousing in Organizations: Issues and Challenges* (pp. 106-118).

www.irma-international.org/chapter/effects-enterprise-resource-planning-erp/18557

INDUSTRY AND PRACTICE: The Dissemination of Good Practices in Database Development Work

Shirley Becker (1998). *Journal of Database Management* (pp. 33-34).

www.irma-international.org/article/industry-practice-dissemination-good-practices/51202

Versioning Approach for Database Evolution

Hassina Bounif (2009). *Handbook of Research on Innovations in Database Technologies and Applications: Current and Future Trends* (pp. 91-101).

www.irma-international.org/chapter/versioning-approach-database-evolution/20692

Model Driven Engineering for Quality of Service Management: A Research Note on the Case of Real-Time Database Management Systems

Salwa M'barek, Leila Baccouche and Henda Ben Ghezala (2016). *Journal of Database Management* (pp. 24-38).

www.irma-international.org/article/model-driven-engineering-for-quality-of-service-management/178634

To Evaluate or Not to Evaluate?: A Two-Process Model of Innovation Adoption Decision Making

Nan (Tina) Wang (2018). *Journal of Database Management* (pp. 42-61).

www.irma-international.org/article/to-evaluate-or-not-to-evaluate/211914