A Nonparametric Outlier Detection for Effectively Discovering Top-N Outliers from Engineering Data

Hongqin Fan\textsuperscript{1}, Osmar R. Zaïane\textsuperscript{2}, Andrew Foss\textsuperscript{2}, and Junfeng Wu\textsuperscript{2}

\textsuperscript{1} Department of Civil Engineering, University of Alberta, Canada
\textsuperscript{2} Department of Computing Science, University of Alberta, Canada

\textbf{Abstract.} We present a novel resolution-based outlier notion and a nonparametric outlier-mining algorithm, which can efficiently identify top listed outliers from a wide variety of datasets. The algorithm generates reasonable outlier results by taking both local and global features of a dataset into consideration. Experiments are conducted using both synthetic datasets and a real life construction equipment dataset from a large building contractor. Comparison with the current outlier mining algorithms indicates that the proposed algorithm is more effective.

1 Introduction

The term “outlier” can refer to any single data point of dubious origin or disproportionate influence. Given a set of observations $X$, an outlier is an observation that is an element of this set but which is inconsistent with the majority of the data or inconsistent with a sub-group of $X$ to which the element is meant to be similar. The above definition has two implications: outlier vis-à-vis the majority; and outlier vis-à-vis a group of neighbours. Whether it is an interesting contaminant or dubious data entry, an outlier is often considered noise, which can have a harmful effect on statistical analysis.

Attempts have been made to remove the noisy data using various outlier mining approaches; in one example, Raz et al. [1] designed an expert system to automatically detect unlikely vehicles and erroneously classified ones from weigh-in-motion data. In contrast to noisy data, some relevant outliers contain important information on system malfunction, mismanagement, or even unpredictable phenomena (environmental or geological disaster), which should be detected for further investigation rather than discarded. In either of these two cases, the inconsistent records should first be identified as much as possible from the dataset. Data validation (range validation, single variate pattern validation etc.) can only filter out a small portion of outliers. Traditional statistical approaches including multivariate outlier detection are not applicable due to their pre-assumption of certain statistical distributions, which may not exist for datasets containing multiple clusters.

Outlier mining techniques in data mining seem to be viable solutions for outlier detection in engineering applications. Some of the popular algorithms include among others a distance-based outlier mining algorithm by Knorr and Ng [2]; a local outlier mining algorithm by Breunig et al. [3] and a connectivity-based mining algorithm by

© Springer-Verlag Berlin Heidelberg 2006
Tang et al. [4]. However these algorithms are not widely accepted in civil engineering disciplines because they do not cater to the special features of engineering datasets, including:

- Current outlier mining algorithms need domain-dependent parameters, but these parameters are not known a priori;
- Current outlier mining algorithms need some parameters, which can only be obtained and tuned through tremendous trial-and-error effort. This is not practical for frequent time-changing applications and thus cannot be an integrated part of a real-time decision support system.
- Current outlier mining algorithms are capable of mining either global or local outliers while the engineering dataset usually contains loosely bounded clusters. It is difficult in this case to differentiate local from global outliers.
- In engineering applications, there exists a need for ranking the top-listed outliers. This is where our major focus is.

In this paper, we present a Resolution-Based outlier (RB-outlier) notion and an associated outlier detection algorithm efficient for engineering applications. The RB-outlier notion is proposed based on a nonparametric clustering algorithm called TURN* by Foss and Zaïane with the same idea of resolution change [5]. The proposed algorithm can detect and rank top-N outliers from any kind of dataset without the need for input parameters.

We also compare RB-outlier with DB-outlier and Local density based outlier (LOF-outlier) mining algorithms using both synthetic datasets and a construction equipment dataset from a large building contractor. Our experimental results show that the RB-outlier mining algorithm generates equivalent or better results than the other two competitive algorithms on all the datasets while benefiting from the absence of input parameters; the RB-outlier results seem to combine the results from both DB-outlier which looks for global outliers and LOF-outlier which looks for local outliers. Analysis on the detected outliers from the equipment datasets shows that these combined results make more sense for engineering datasets.

2 Related Work

Hawkins defines an outlier as “an observation which deviates so much from other observations as to arouse suspicions that it was generated by a different mechanism” [6]. Traditionally outlier detection in engineering disciplines depends on statistical approaches. After fitting a data series into a bell-shaped statistical distribution, those data points located far away from the mean (e.g. 3 standard deviations) are deemed outliers; multivariate outlier detection techniques can help identify outliers within a multivariate dataset. Other commonly used techniques are quartile methods, and visualization methods using scatter plot, etc. With regard to the data complexity and sheer data volume in engineering systems, outlier detection using statistical approaches is very inefficient and even impractical due to their limitations such as the difficulties with handling higher dimensional data, and the necessary assumption of distributions.