Data Min Knowl Disc (2010) 20:189–190
DOI 10.1007/s10618-009-0163-0

Outlier detection special issue

Sanjay Chawla · David Hand · Vasant Dhar

Received: 23 December 2009 / Accepted: 28 December 2009 / Published online: 29 January 2010

An outlier is an unexpected event or entity. For example, many experts view the Great Financial Crash of 2008 as an outlier event which has triggered a reappraisal of mainstream economic and financial models which define the "expected." The objective of Outlier Detection in Data Mining is in similar vein—outliers often embody new information, which is often hard to explain in the context of existing knowledge and results in a re-evaluation of what is known.

An important theme in Statistics/Machine Learning/Data Mining is the use of data to study the "norm" or "expected" behaviour of the underlying phenomenon (which generated the data). The presence of outliers often distorts the understanding of the norm and has given rise to a set of techniques, often called robust statistics, which discount the effect of outliers. A canonical example is the use of the median (which is less sensitive to outliers) as opposed to the mean (which is extremely sensitive to outliers) to characterize average behaviour.

In Data Mining, an outlier is a primary object of study which can potentially lead to the discovery of new "knowledge." Thus the emphasis has been on the design of algorithms to find outliers in complex scenarios while relaxing as many assumptions on the underlying data generating model as possible. Or more simply, the focus in Data Mining is on non-parametric (and semi-parametric) outlier detection techniques. Here is a simple example:

Let D be a multi-variate data set and the objective it to discover whether there are any outliers in D. If we assume that D was generated from a multi-dimensional Normal distribution then it is well known that the Mahalanobis distance from data points to the

Responsible Editor: Geoffrey Webb.

S. Chawla (✉) · D. Hand · V. Dhar
University of Sydney, Sydney, Australia
e-mail: chawla@it.usyd.edu.au

mean of D follows a Chi-Square distribution. Outliers are data points which reside in the tail of the Chi-Square distribution. Thus we arrive at an *analytical* solution to the problem. In Data Mining, we abstain from making assumptions on the data generating process. However, we do assume the existence of a similarity measure (often domain dependent) which can be used to quantify the "similiarity" between any pair of data points in D. Outliers are those points which are least similar to their neighbors. Thus for each data point we have to search for its neighbors. This results in *a combinatorial* problem.

This special issue, consisting of four papers, has a combinatorial and algorithmic bias.

Wu, Xiong and Chen [WGC] in "COG: Local Decomposition of Rare Class Analysis" propose an innovative solution to the imbalanced classification problem which can be viewed as a supervised learning approach for outlier detection. It is well known that the performance of standard classification techniques deteriorates in the presence of imbalanced data. The method proposed by WGC first uses clustering to breakdown the majority classes into smaller classes and then uses standard classification techniques to learn the different classes (including the rare or minority class). A detailed analytical and experimental analysis is carried out to test the efficacy of their approach.

In "Spatial Neighborhood based Anomaly Detection in Sensor Datasets," Janeja, Adam, Atluri and Vaidya, propose the use of micro-neighborhoods to capture spatial autocorrelation and heterogeneity between spatial entities. The similarity relationship defined between entities is then used to search for outliers in large databases which have an explicit spatial and temporal component. They test their approach in diverse situations like water and highway traffic monitoring.

Koufakou and Georgiopolous in "A Fast Outlier Detection Strategy for Distributed High-Dimensional Data Sets with Mixed Attributes," present algorithms for the detecting outliers in situations where data is distributed across many sources and comprises both categorical and continuous attributes. Since they also assume data is high dimensional, they choose to use to cosine similarity for continuous attributes. A new similarity measure is also proposed for categorical data.

In "Distance-based Outlier Queries in Data Streams: the Novel task and Algorithms," Fabrizio and Fabio present innovative approaches to mine for outliers in settings where data is continuously arriving and the system has limited available memory to process for outliers. They present three algorithms which neatly illustrate the trade-off between accuracy, time and space.

Selection process

The special issue received over twenty five papers and after a careful and extensive peer-review process, four papers were selected which cover diverse aspects of outlier detection. We hope the readers will find these papers beneficial and that this special issue will catalyse further interest in this important area of Data Mining.