An Adaptive Outlier Detection Technique for Data Streams*

Shiblee Sadik and Le Gruenwald

School of Computer Science, University of Oklahoma, Norman, OK 73071

Abstract. This work presents an adaptive outlier detection technique for data streams, called Automatic Outlier Detection for Data Streams (A-ODDS), which identifies outliers with respect to all the received data points (global context) as well as temporally close data points (local context) where local context are selected based on time and change of data distribution.

1 Introduction

An outlier is a data point which is significantly different from other data points [1]. Although outliers are interesting to the user, a handful of techniques are available for data streams, which are adopted from existing outlier detection techniques for regular data with ad-hoc modifications. A number of those techniques use sliding window and detect outliers inside the window [2]; but an outlier for a particular window may appear as an inlier in another window; hence the notion of outlier in a data stream window is not very concrete. Auto-regression based techniques construct a model and compute a metric for each data point [3] where a data point is an outlier if the corresponding metric is beyond a certain cut-off limit. However finding a proper auto-regression model and cut-off limit is a not a trivial task and requires expert knowledge. Statistics based techniques [1] assume a fixed data distribution while data streams have varying distribution. In this work we present A-ODDS to detect outliers based on the deviations of a data point with respect to global and local contexts.

2 The Proposed Technique: A-ODDS and Experimental Results

Our approach is based on two deviation factors for the global and local contexts, called Global Deviation Factor (GDF) and Local Deviation Factor (LDF), respectively. GDF represents the deviation of a data point with respect to the entire history data points; and LDF represents the deviation of a data point with respect to the recent data points; both deviation factors are calculated from neighbor density.

GDF of a data point is the relative distance from the average neighbor density of the entire history data points to its neighbor density; and LDF of a data point is the relative distance from the average neighbor density of the recent data points to its neighbor density. A data point is identified as an outlier if either its GDF or LDF goes beyond three standard deviations away from its respective average. The choice of

* This work has been supported in part by the NASA under the grants No. NNG05GA30G.

J. Bayard Cushing, J. French, and S. Bowers (Eds.): SSDBM 2011, LNCS 6809, pp. 596–597, 2011.

three standard deviation dispersion ensures a significant dispersion of a data point from other data points [1] and does not require the user to select cut-off limits.

Our local context selection scheme for LDF is based on two intuitive criteria: first, data points in local context have to be temporally close and second, they have to follow similar distribution. We choose data points in between two consecutive concept drifts as local context as they are close temporally and expected to follow similar distribution; hence LDF finds outliers non-conformist to the recent trend. GDF and LDF use the dynamically adaptive data distribution function for neighbor density computation that we presented in [5].

We conducted simulation experiments using a real dataset collected from California Irrigation Management [4] to compare A-ODDS with the three existing algorithms: auto-regression based algorithm ART [3], sliding window based algorithm ODTS [2] and distance-based outlier detection algorithm DBOD-DS [5], in terms of outlier detection accuracy (Jaccard Coefficient (JC) [2]) and execution time. On average, as shown in Table 1, A-ODDS gives the best accuracy among all the techniques; however, when measuring execution time, while A-ODDS requires less time than DBOS-DS, it takes more time than ART and ODTS.

Table 1. Average JC and Execution Time

	Average accuracy (JC)	**Average execution time (ms)**
A-ODDS	0.7095	1.3656
DBOD-DS	0.1585	1.8485
ODTS	0.1467	0.0405
ART	0.1373	0.3005

3 Conclusions and Future Work

This paper presents an overview of A-ODDS and its accuracy and efficacy compared to existing algorithms. For future work, we will perform extensive empirical studies and extend it to multi-dimensional and multiple heterogeneous data streams.

References

1. Barnett, V., Lewis, T.: Outliers in Statistical Data. Wiley Series in Probability and Mathematical Statistics. John Wiley & Sons Inc., Chichester (1994)
2. Basu, S., Meckesheimer, M.: Automatic outlier detection for time series: an application to sensor data. Knowledge Information System (2007)
3. Curiac, D., Banias, O., Dragan, F., Volosencu, C., Dranga, O.: Malicious Node Detection in Wireless Sensor Networks Using an Autoregression Technique. In: ICNS 2007 (2007)
4. California Irrigation Management Information System. web-link, `http://wwwcimis.water.ca.gov/cimis/welcome.jsp` (accessed January 2010)
5. Sadik, S., Gruenwald, L.: DBOD-DS: Distance Based Outlier Detection for Data Streams. In: Bringas, P.G., Hameurlain, A., Quirchmayr, G. (eds.) DEXA 2010. LNCS, vol. 6261, pp. 122–136. Springer, Heidelberg (2010)