Anomalies in multidimensional contexts

N. Dunstan, I. Despi & C. Watson
School of Science and Technology, University of New England, Australia

Abstract

This paper investigates the problem of presenting anomalies in a multidimensional data set. In such a data set, some dimensions may be merely descriptive, while others represent measures and attribute values used to determine whether the data is anomalous. A data cube of the descriptive dimensions is used as a data structure to partition the data set into subgroups at each node, or context. It is shown that it is possible for a datum to be anomalous in more than one context. Previous work has dealt with this problem by embedding exception indicators in the data cube. Since the data cube is potentially large and anomalies are rare, searching for anomalies is inconvenient. Instead, it is proposed to construct a report for each anomaly that shows its status in each possible context. This results in a direct presentation of anomalous data.
Keywords: anomalies, outliers, exception indicators, data cubes, decision support, discovery-driven exploration, knowledge management.

1 Introduction

Transaction processing systems generate high volumes of data, often including descriptive details of the transaction. Transaction records are thus multidimensional, with each dimension possibly contributing to useful information that may be discovered on analysis. An important task of transaction data set analysis is to uncover and investigate anomalies. Applications include data cleansing by removing outliers in preparation for data mining activities (Han and Kamber [1]), and fraud detection (Zakai and Akira [10]).

There have been many algorithms developed to find anomalies (or outliers) in multidimensional space (Knorr and Ng [6], Ramaswamy *et al.* [8], Ceglar *et al.* [4], Chaudhary *et al.* [3]). Of particular relevance to this paper is the approach of Aggrawal and Yu [2] which searches for anomalies, not just in the full dimensional


WIT Transactions on Information and Communication Technologies, Vol 42, © 2009 WIT Press
www.witpress.com, ISSN 1743-3517 (on-line)
doi:10.2495/DATA090181

space, but also in subdimensional spaces (or projections). Sarawagi *et al.* [9] developed a data cube model with cells annotated with exception indicators to guide users to the discovery of anomalies. Li and Han [7] also used the data cube concept in an algorithm to find anomalies in time series data. A data cube (Gray *et al.* [5]) is a lattice of summarization tables of different combinations of dimensions and at different levels of detail.

In this paper, a reporting method for anomalies in multidimensional transaction records is proposed. It uses the data cube framework as a means of partitioning the data set into subgroups, or contexts. It is shown that a record may be anomalous in more than one context, and that data cube cells (representing subgroups of transactions) may be anomalous in higher-level contexts. Since data cubes may be very large and anomalies very rare, our presentation of anomalies to users is independent of the summarization table presentation, while still retaining the hierarchical information inherent in the data cube structure.

2 Multidimensional data and data cubes

Transaction records can be represented in the form of a table, with each row indicating the attribute values of each record. An example is shown in Figure 1, where records indicate sales amounts and number of items bought, as well as information about the place and day of purchase, and the buyer. These records may be said to have seven dimensions, since attribute values are recorded for each of seven different attributes, or column headings.

The data cube is constructed using dimensions that are of interest to the user. There are 2^n nodes (or cuboids) in the data cube, where n is the number of dimensions chosen. To illustrate using the example of Figure 1, a data cube using the dimensions `Branch`, `DayofWeek`, `Gender` and `Member`, is shown in Figure 2.

Each node of the data cube represents a table of aggregate values. Aggregations can involve operators such as `sum`, `count`, `minimum` and `maximum`, and these are chosen by the user. Each table will contain a row (or cell) for each possible combination of attribute values. A table can be described using the `SQL Group By` operator. For example, if the original table of Figure 1 is called `t1`, then the table `BranchDayofWeek` (`BD` in Figure 2) would be defined as:

```
create table BranchDayofWeek
select Branch, DayofWeek,
sum(Amount) as Sum, count(Amount) as Count
from t1 group by Branch, DayofWeek;
```

Assuming that `Branch` has only two valid values, `armidale` and `guyra`, and that `DayofWeek` has seven, then there will be fourteen rows in the table, as illustrated in Figure 3. The `Apex` table is a one row aggregation of all the transactions in the data set.

TID	Branch	DayofWeek	Gender	Member	Amount	Items
T1	armidale	tuesday	female	no	60	2
T2	armidale	tuesday	female	no	50	1
T3	guyra	wednesday	male	yes	70	2
...	...	...	...	...	...	...

Figure 1: Transaction records with multidimensions.

A cell can be uniquely referenced using a vector notation. For example,

$$(armidale, monday, *, *)$$

is the first cell of the cuboid `BranchDayofWeek`. The $*$ symbols indicate that dimensions `Gender` and `Member` are not specified.

Clearly, each transaction record will belong to one cell in each cuboid. For example, `T1` from Figure 1 belongs to:

$(armidale, tuesday, female, no)$ in `BranchDayofWeekGenderMember`,
$(armidale, tuesday, female, *)$ in `BranchDayofWeekGender`, and
$(armidale, tuesday, *, *)$ in `BranchDayofWeek`.

Each cuboid represents a *context* in which the transaction data set is partitioned into different subgroups.

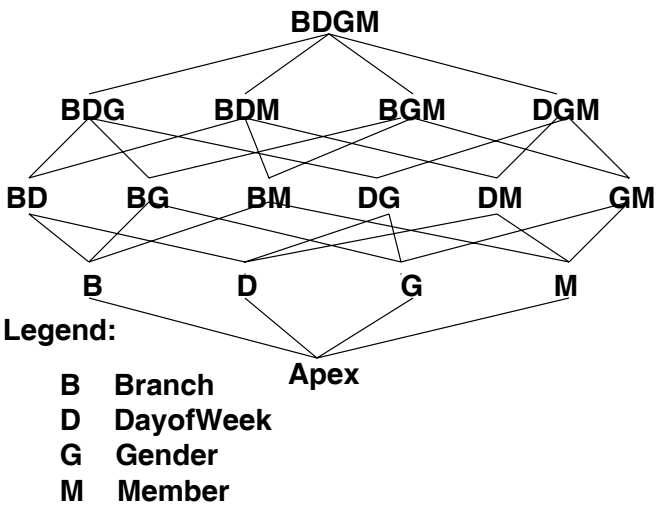

Figure 2: Data cube.

Branch	DayofWeek	Sum	Count
armidale	monday	...	...
	tuesday	...	...
	...	...	...
	sunday	...	...
guyra	monday	...	...
	...	...	...

Figure 3: Table Branch DayofWeek.

3 Anomalies in data cube contexts

3.1 Anomalous Transactions

The simple data set example in Figure 4 is used to show that a transaction may be anomalous in more than one context. It has just two dimensions (A and B) used to form the data cube. And only one dimension (m) used to calculate whether or not the transaction is an anomaly. In this simple example, a transaction is defined to be an anomaly in a subgroup if its absolute distance from the mean is more than 1.5 standard deviations. That is

$$|MEAN - m| - 1.5 \times STD$$

The data cube is thus just four tables. They are shown in Figure 5.

TID	A	B	m
D1	a1	b1	40
D2	a1	b1	50
D3	a1	b1	55
D4	a1	b1	80
D5	a2	b1	60
D6	a2	b2	80
D7	a2	b2	80
D8	a2	b2	55
D9	a2	b2	70
D10	a3	b2	90

Figure 4: Transaction data set.

AB	A	B	count	sum m	MEAN	STD
	a1	b1	4	225	56.3	14.7
	a1	b2	0	0	0	0
	a2	b1	1	60	60.0	0
	a2	b2	4	285	71.3	10.2
	a3	b1	0	0	0	0
	a3	b2	1	90	90	0

A	A	B	count	sum m	MEAN	STD
	a1	*	4	225	56.3	14.7
	a2	*	5	345	69.0	10.2
	a3	*	1	60	60.0	0

B	A	B	count	sum m	MEAN	STD
	*	b1	5	285	57.0	13.3
	*	b2	5	375	75.0	11.8

Apex	A	B	count	sum m	MEAN	STD
	*	*	10	660	66.0	15.5

Figure 5: Transaction data cube.

Context / TID	AB	A	B	Apex
D1	–5.8	–5.8	–2.9	2.8
D4	1.6	1.6	3.1	–9.2
D8	0.9	–1.3	2.3	–12.2
D10	0	0	–2.7	0.8

Figure 6: Transaction anomalies by contexts.

In a real-world transaction data set, the data cube would be large and anomalies rare. A concise presentation of anomalies in the contexts defined by the data cube is a table of anomalous transactions by contexts (cuboids). For the example data set of Figure 4, the anomaly table is shown in Figure 6, where a positive number indicates an anomaly. It should be noted that some subgroups have too few transactions to have anomalies under this definition.

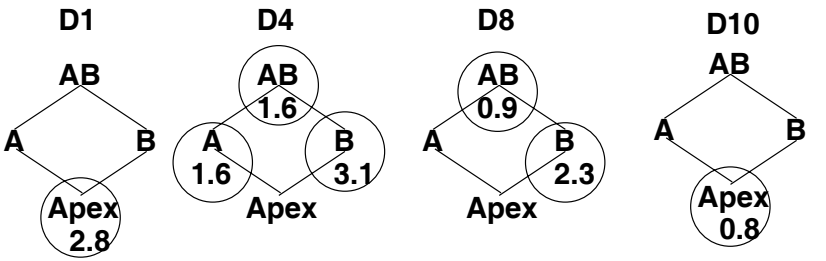

Figure 7: Anomaly lattices.

3.2 Anomalous subgroups

It is also possible that cells, or subgroups of transactions might be identified as anomalous. For example, a low, or zero `count` for a subgroup for sales in a particular day. It is for this reason that a detection algorithm that considers projections, as well as the full dimensional space, is desirable. Cells are also parts of higher subgroups in the data cube, and thus, can also be anomalous in contexts. This may be represented in tables in a similar way to that of Figure 6.

4 Lattice representation

Tables such as Figure 6 are a concise, textual way of presenting information about anomalies in different contexts. The contexts are derived from the data cube. A graphical way of presenting the same information is to use the data cube lattice, annotated with the anomaly measure. This is illustrated in Figure 7. Under the example definition, transactions `D1`, `D4`, `D8` and `D10` are anomalies. `D1` and `D10` are anomalies only in the broadest context (the `Apex`). `D4` and `D8` are anomalies in three and two contexts respectively.

5 Conclusions

The detection of anomalies in transaction data sets is important in many fields of business, security and data mining. A data cube is a decision support tool that is constructed using domain knowledge of data dimensions that are important in describing the data. This paper uses the data cube structure to define contexts and shows how individual transactions can be anomalous in different user-defined contexts. Anomaly tables and lattices are concise ways to present anomaly information in multidimensional data sets.

References

[1] Han, J. and Kamber, M., *Data mining: concepts and techniques*, Morgan Kaufman, 2nd edition, 2006.

[2] Aggrawal, C.C., and Yu, P.S., An effective and efficient algorithm for high-dimensional outlier detection, *The VLDB Journal*, **14**, pp. 211–221, 2005.
[3] Chaudhary, A., Szalay, S. and Moore, A., Very Fast Outlier Detection in Large Multidimensional Data Sets, *Data Mining and Knowledge Discovery*, ACM Press, 2002.
[4] Ceglar, A., Roddick, J.F., and Powers, D.M.W., CURIO: A fast outlier and outlier cluster detection algorithm for large datasets, *Proceedings of the 2nd International Workshop on Integrating Artificial Intelligence and Data Mining*, Gould Coast, pp. 39–48, 2007.
[5] Gray, J., Chaudrhuri, A., Bosworth, A., Layman, A., Reichart, D., Venkatrao, M., Pellow, F. and Pirahesh, H., Data cube: a relational aggregation operator generalizing group-by, cross-tab and sub-totals, *Data Mining and Knowledge Discovery*, **1**, pp. 29–54, 1997.
[6] Knorr, E. and Ng, R., Algorithms for mining distance-based outliers in large datasets, *Proceedings of the VLDB Conference*, New York, September, 1998.
[7] Li, X. and Han, J., Mining approximate top-k subspace anomalies in multi-dimensional times-series data, *VLDB'07*, Vienna, September, pp. 447–458, 2007.
[8] Ramaswamy, S., Rastogi, R. and Shim, K., Efficient algorithms for mining outliers from large data sets, *ACM SIGMOD International Conference on Management of Data*, Dallas, pp. 427–438, 2000.
[9] Sarawagi, S., Aggrawal, R., and Megiddo, N., Discovery-driven exploration of OLAP data cubes, *Proceedings of the International Conference of Extended Database Technology*, Valencia, March, 1998.
[10] Zakia, F. and Akira, M., *Joho Shori Gakkai Zenkoku Taikai Koen Ronbunshu*, **68**(3), pp. 279–280, 2006.