Mining Multidimensional Contextual Outliers from Categorical Relational Data

Guanting Tang  
Simon Fraser University  
gta9@cs.sfu.ca

Jian Pei  
Simon Fraser University  
jpei@cs.sfu.ca

James Bailey  
The University of Melbourne  
baileyj@unimelb.edu.au

Guozhu Dong  
Wright State University  
guozhu.dong@wright.edu

ABSTRACT
A wide range of methods have been proposed for detecting different types of outliers in full space and subspaces. However, the interpretability of outliers, that is, explaining in what ways and to what extent an object is an outlier, remains a critical open issue. In this paper, we develop a notion of contextual outliers on categorical data. Intuitively, a contextual outlier is a small group of objects that share strong similarity with a significantly larger reference group of objects on some attributes, but deviate dramatically on some other attributes. We develop a detection algorithm, and conduct experiments to evaluate our approach.

Keywords  
Outlier detection, context, categorical data

1. INTRODUCTION
Outlier detection is an important data mining task. A wide range of methods have been proposed to detect different types of outliers on various kinds of data. Interpretability of outliers, however, remains a serious concern. More often than not, an analyst may want to see not only the outliers detected, but also insightful explanations about the outliers. Particularly, an analyst may want to know, for an outlier, a reference group of objects from which the outlier deviates in some aspects and shares similarity with in some other aspects, and a set of features manifesting the outlier’s unusual/deviating behavior, the outlier degree, and the other similar outliers sharing the same context. Such contextual information can help an analyst to better understand and investigate individual outliers and propose action plans suitable for such outliers. We argue that the contextual information about outliers should be an integral component in the outlier detection process. Unfortunately, most of the existing outlier detection methods do not provide rich and detailed contextual information for outlier analysis.

In this paper, we tackle the problem of contextual outlier detection on categorical data, and make two main contributions. First, we develop a notion of multidimensional contextual outliers to model the context of an outlier. Intuitively, a contextual outlier is a small group of objects that share similarity on some attributes with a significantly larger reference group of objects, but deviate dramatically from the reference group on some other attributes. An example is: “Among the computer science senior undergraduate students at University X, a small group of 3 students not enrolled in the data structure course is an outlier against the reference group of 128 students enrolled in the course.” In contextual outlier detection, we identify not only the outliers, but also their associated contextual information including (1) comparing to what reference group of objects the detected object(s) is/are an outlier; (2) the attributes defining the unusual behavior of the outlier(s) compared against the reference group; (3) the population of similar outliers sharing the same context; and (4) the outlier degree, which measures the population ratio between the reference normal group and the outlier group. Second, we design a simple yet effective contextual outlier detection algorithm that leverages the state-of-the-art data cube computation techniques. The focus of our method is to find outliers together with their contextual information. We conduct experiments to evaluate the feasibility and usefulness of our approach.

The rest of the paper is organized as follows. We propose the notion of contextual outliers in Section 2, and develop a detection algorithm in Section 3. We evaluate our approach in Section 4. We review related work in Section 5, and conclude the paper in Section 6.

2. CONTEXTUAL OUTLIERS
In this paper, we consider outlier detection on multidimensional categorical data. Specifically, we consider a base table $T(A_1, \ldots, A_n)$, where $A_1, \ldots, A_n$ are categorical at-
tributes with finite domains. We assume that each object is represented by a tuple in the base table and is associated with an identifier tid, which is used as a reference to the object only, and does not carry any other meaning. For an object \( t \in T \), denote by \( t.A_i \) and \( t.tid \) the value of \( t \) on attribute \( A_i \) (1 \( \leq i \leq n \)) and the identifier of \( t \), respectively.

A subspace is a subset of attributes. In order to summarize a group of objects, we add a wildcard meta-symbol \(*\) to the domain of every attribute \( A_i \) (1 \( \leq i \leq n \)). Symbol \(*\) matches any possible values in the domain. A group-by tuple (or group for short) is a tuple \( g = (g.A_1, \ldots, g.A_n) \) such that \( g.A_i \) takes either a value in the domain of \( A_i \) or meta-symbol \(*\). The cover of \( g \) is the set of objects in \( T \) matching \( g \), that is, \( \text{cov}(g) = \{ t \in T \mid \forall i, g.A_i \neq * \} \). The set space\(^{(i)}\)(\( g \)) = \{ \( A_i \mid 1 \leq i \leq n, g.A_i \neq * \} \) is called the subspace of \( g \), and the set \( \text{avs}(g) = \{ A_i = g.A_i \mid 1 \leq i \leq n, g.A_i \neq * \} \) is called the non-* attribute-value set (AVS) for short) of group \( g \). For an AVS \( V \), we overload the operator \( \text{space}() \) by defining \( \text{space}(V) = \{ A_i \mid A_i \in V \} \). Thus, \( \text{space}(\text{av}(g)) = \text{space}(g) \).

For two distinct groups \( g_1 \) and \( g_2 \), \( g_1 \) is an ancestor of \( g_2 \), and \( g_2 \) a descendant of \( g_1 \), denoted by \( g_1 \supset g_2 \), if \( \text{avs}(g_1) \subset \text{avs}(g_2) \), that is, for every attribute \( A_i \) (1 \( \leq i \leq n \)) such that \( g_1.A_i \neq * \), we have \( g_2.A_i = g_1.A_i \). We write \( g_1 \supset g_2 \) if \( g_1 \supset g_2 \) or \( g_2 \supset g_1 \).

**Property 1 (Monotonicity).** For two groups \( g_1 \) and \( g_2 \) such that \( g_1 \supset g_2 \), \( \text{cov}(g_1) \supset \text{cov}(g_2) \).

Intuitively, for a group of outlier objects, the contextual information consists of a group of reference objects that manifest the outlier group in a subspace. The comparison of the two groups in population size is also included.

**Definition 1 (Contextual outlier).** Let \( T \) be a base table, and \( r, o \) be two groups such that \( \text{space}(r) = \text{space}(o) \neq \emptyset \). Given an outlier degree threshold \( \Delta > 1 \), the pair \((r, o)\) is a contextual outlier if the outlier degree \( \text{deg}(r, o) = \frac{|\text{cov}(r)|}{|\text{cov}(o)|} \geq \Delta \). We call \( r \) the reference group, \( o \) the outlier group, \( \text{out}(r, o) = \text{space}(r) - \text{space}(\text{cond}(r, o)) \) the outlier subspace, and \( \text{cond}(r, o) = \text{avs}(r) \cap \text{avs}(o) \) the shared AVS. It is possible that \( \text{cond}(r, o) \) is empty.

The shared AVS \( \text{cond}(r, o) \) provides a context subspace for the outlier analysis about \( o \). The objects in groups \( o \) and \( r \) belong to the same context subspace, that is, they take the same values on those attributes that occur in \( \text{cond}(r, o) \). If \( \text{cond}(r, o) = \emptyset \), \( r \) and \( o \) do not share any common features. In such a special case, \( o \) is a global outlier that is small in population and different from a large reference group \( r \) in space \( \text{space}(o) = \text{space}(r) \).

The reference group \( r \) indicates the normal or dominating objects to which \( o \) is compared. The outlier group \( o \) and the outlier subspace \( \text{out}(r, o) \) indicate the outlier objects \( \text{cov}(o) \) and the attributes that manifest the deviation of \( o \) from \( r \). The outlier degree measures how exceptional the group \( o \) is when compared to \( r \). The larger the outlier degree is, the more outlying \( o \) is.

### 3. Detection Algorithms

In this section, we develop an algorithm for contextual outlier detection. We observe that group-bys are essential units in both data cube computation and contextual outlier analysis, so we can exploit the state-of-the-art data cube techniques in detecting contextual outliers.

**Lemma 1 (Non-closure Attributes).** For two contextual outliers \((r_1, o_1)\) and \((r_2, o_2)\) in a base table \( T \), if \( r_1 \supset r_2, o_1 \supset o_2, \text{cov}(r_1) = \text{cov}(r_2), \) and \( \text{cov}(o_1) = \text{cov}(o_2), \) then \( \text{deg}(r_1, o_1) = \text{deg}(r_2, o_2) \).

In the two contextual outliers \((r_1, o_1)\) and \((r_2, o_2)\) in Lemma 1, the two groups \( r_1 \) and \( r_2 \) capture the same set of objects. Hence \((r_1, o_1)\) is redundant given \((r_2, o_2)\) or vice versa. Since \( r_1 \supset r_2 \), \( r_2 \) contains some extra attributes in addition to those in \( r_1 \). Hence \( r_2 \) is more informative and descriptive than \( r_1 \) as a reference group. It is better to include \((r_2, o_2)\) for outlier analysis.

**Definition 2 (Closure group/outlier).** Given a base table \( T \), a group \( g \) is a closure group if for any descendant group \( g' \subset g \), \( \text{cov}(g') \subset \text{cov}(g) \). \((r, o)\) is called a closure outlier if there does not exist another contextual outlier \((r', o')\) such that \( r' \supset r, o' \supset o, \text{cov}(r) = \text{cov}(r'), \) and \( \text{cov}(o) = \text{cov}(o') \).

**Theorem 1 (Closure group/outlier).** Contextual outlier \((r, o)\) is a closure outlier if and only if either \( r \) or \( o \) is a closure group.

Since every closure contextual outlier must have either the reference group or the outlier group as a closure group, we can find all closure groups in the base table first, and then use the closure groups to find contextual outliers. Finding closure groups and closure patterns has been well studied in frequent pattern mining [12, 16] and data cube computation [8]. Given a base table \( T \), we can adopt a state-of-the-art algorithm, such as the DFS algorithm in [8], to find all closure groups.

Algorithm 1 presents the pseudocode of our detection method, COD (for Contextual Outlier Detection). For each closure group \( o \), we consider all the other closure groups \( r \) such that \( (r, o) \) is a contextual outlier. Obviously, \( \text{cov}(o) \) cannot be larger than \( \frac{1}{l} \), where \( l \) is the largest cover size among all closure groups (calculated in Line 1). For each of such closure groups \( o \), we iterate over all the other closure groups \( r \) such that \( \text{cov}(r) \geq \Delta \text{cov}(o) \) (the inner loop, Lines 5-7). The iteration continues until all closure groups that may be outlier groups are examined.

### 4. Experimental Results

In this section, we report our empirical evaluation of COD using real data sets. All experiments were conducted on a PC computer with an Intel Core Duo E8400 3.0 GHz CPU and 4 GB main memory, running the Microsoft Windows 7 operating system. The algorithms were implemented in C++ using Microsoft Visual Studio 2010.
Algorithm 1 COD: the contextual outlier detection algorithm.

Require: $G$: the complete set of closure groups; $\Delta$: the outlier degree threshold

Ensure: the complete set of contextual outliers

1: let $l = \max_{g \in G} \{|\text{cov}(g)|\}$;
2: let $O$ be the set of contextual outliers; set $O = \emptyset$;
3: for each closure group $g$ such that $|\text{cov}(g)| \leq \frac{1}{\Delta} l$ do
4: for each closure group $r$ such that (1) $|\text{cov}(r)| \geq \Delta |\text{cov}(g)|$; and (2) $\text{space}(r) \subseteq \text{space}(g)$ or $\text{space}(r) \supseteq \text{space}(g)$ do
5: $O = O \cup (r', g')$;
6: end for
7: end for
8: return $O$

Table 1: The statistics of the data sets.

<table>
<thead>
<tr>
<th>Data Set</th>
<th>Solar-flare</th>
<th>Tic-tac-toe</th>
<th>Credit-approval</th>
<th>Hayes-roth</th>
</tr>
</thead>
<tbody>
<tr>
<td># obj.</td>
<td>1,389</td>
<td>958</td>
<td>600</td>
<td>160</td>
</tr>
<tr>
<td># attr.</td>
<td>10</td>
<td>9</td>
<td>8</td>
<td>4</td>
</tr>
<tr>
<td># closure grp.</td>
<td>7,770</td>
<td>42,711</td>
<td>5,707</td>
<td>277</td>
</tr>
<tr>
<td>QC time (s)</td>
<td>2.136</td>
<td>12.903</td>
<td>1.446</td>
<td>0.047</td>
</tr>
</tbody>
</table>

Note: QC time refers to the time used to find all closure groups, that is, $|G|$ in algorithm 1.

We cannot identify any existing method that solves the exact same problem. The focus of our method is to find outliers with contextual information. Consequently, this paper does not intend to compete with the existing methods.

We use categorical data sets from the UCI repository [6]. We report the results on four data sets: solar-flare, tic-tac-toe, credit-approval, and hayes-roth. Some statistics of the data sets are summarized in Table 1. We evaluate COD in two aspects: effectiveness and efficiency.

4.1 Case Studies

We demonstrate the effectiveness of context outlier detection using case studies on the data set hayes-roth.

The hayes-roth data set records the information about 160 people on four attributes [1, 6]. The first attribute, hobby, takes values uniformly at random [6] and we thus ignore it in our analysis, that is, all groups take value * on the attribute. The description for the other three attributes are adopted from [1]: the second attribute, age, takes values in \{30, 40, 50, r > 0\} (the meaning of value “r > 0” is unspecified in [1]); attribute education takes values in {junior-high, high-school, trade-school, college}; and the last attribute, marital-status, takes values in {single, married, divorced, widowed}. Table 2 shows some interesting contextual outliers with respect to $\Delta = 5$.

In Table 2, outliers $c_1$ and $c_2$ share the same reference group. The reference group consists of 34 people whose marital-status is “single” and who have high-school degrees. Outlier group $c_1$ is a collection of 6 “divorced” college graduates and outlier group $c_2$ is a collection of 6 “single” trade school graduates. Outlier $c_3$ is interesting: among people who are divorced, those who are college graduates are outliers compared to those with high school degrees. $c_3$ shows that, among people who are 50-years old, the 2 with college degrees are outliers compared to the 16 with high school degrees. Another interesting outlier is $c_{10}$: in the age group of 30, the 4 people widowed are outliers compared to the 34 people married. Please note that, in the whole data set, there are 59 of high-school, 59 of junior-school, 29 of tradeschool and 13 of college graduates. Given $\Delta = 5$, those of trade-school and college graduates are not outliers comparing to those of high-school and junior-school. The outliers can only be explained well using the contextual information.

4.2 Efficiency

COD takes only one parameter, the outlier degree threshold $\Delta$. Figure 1 shows the runtime of COD on the four real data sets with respect to various $\Delta$ thresholds. The closure group computation time is reported in Table 1, and is not included.

5. RELATED WORK

To the best of our knowledge, only very few existing studies consider context in outlier detection. Song et al. [14] proposed the notion of conditional outliers to model the outliers manifested by a set of behavioral attributes (e.g. temperature) conditionally depending on a set of contextual attributes (e.g. longitude and latitude). The behavioral attributes and the contextual attributes are pre-defined. Our contextual outlier model automatically identifies shared AVSs. Moreover, reference groups are not modeled in [14]. A mixture model is used in [14] to mine conditional outliers, which is infeasible in our model since here the subspaces are not pre-defined and change from one outlier to another. Valko et al. [15] detected conditional anomalies using a training set of labeled examples with possible label noise, which is different from our work that no labeled data is assumed.

Kriegel et al. [7] proposed a method that detects an outlier with reference to the axis parallel subspace spanned by its neighbors. Müller et al. [11] proposed a technique for ranking outliers based on their degree of deviation in different subspace projections. While these studies also focus on subspace context for outliers, there are two key differences from our work. First, these studies focus on continuous datasets,
### Table 2: Some contextual outliers on data set hayes-roth. The underlined attributes indicate the shared AVSs.

| Outlier-id | Reference group $r$ | Outlier group $o$ | $deg(r, o) = \frac{\text{outlier degree}}{|r|}$ |
|-----------|---------------------|-------------------|---------------------------------------------|
| c1        | (*, *, high-school, single) | (*, *, high-school, divorced) | 5.7 = 34/6 |
| c2        | (*, *, high-school, single) | (*, *, trade-school, single) | 5.7 = 34/6 |
| c3        | (*, *, high-school, single) | (*, *, high-school, widowed) | 8.5 = 34/4 |
| c4        | (*, *, trade-school, married) | (*, *, trade-school, widowed) | 8.0 = 16/2 |
| c5        | (*, *, junior-high, divorced) | (*, *, college, divorced) | 8.0 = 16/2 |
| c6        | (*, college, *, *) | (*, college, *, *) | 8.5 = 34/4 |
| c7        | (*, college, *, *) | (*, junior-high, *, *) | 8.5 = 34/4 |
| c8        | (*, junior-high, *, *) | (*, junior-high, *, *) | 8.5 = 34/4 |
| c9        | (*, *, high-school, *) | (*, college, *, *) | 8.0 = 16/2 |
| c10       | (*, *, *, married) | (*, college, widowed) | 8.5 = 34/4 |

7. REFERENCES

while our focus is on categorical relational data. Second, our work proposes techniques for concise descriptions of sets of outliers. We also provide contextual descriptions for the outliers that are detected.

Recently, Smet and Vreeken [13] developed an outlier detection method $OC^3$, which assumes that outliers are generated by a distribution different from that generates the normal objects, and uses minimum description length (MDL) to detect outliers. Again, the notions of context, reference groups and outlier groups are not modeled simultaneously in $OC^3$. Angiulli et al. [2] studied a related by orthogonal problem. Given a multidimensional database and a query object in the database, find the top- $k$ subset of attributes that the query object receives the highest outlier score. Their method does not find outliers directly. Moreover, it finds subspaces but does not find reference groups in outlier explanation.

A contextual outlier $(r, o)$ identified in our method can be written as a pair of rules: $\text{cond}(r, o) \Rightarrow \text{avs}(r) - \text{cond}(r, o)$ for the reference group, and $\text{cond}(r, o) \Rightarrow \text{avs}(o) - \text{cond}(r, o)$ for the outlier group. There are a number of methods using rules in outlier detection [5, 3, 10, 17]. Our method differs from the existing methods in several aspects. First, most of the existing rule-based methods focus on detecting individual outliers, and may not be able to identify outlier groups and measure the outlyingness accordingly. Second, many existing rule-based methods use rules to model only the normal objects or strong associations. Outliers are individual objects that do not follow those rules. Those methods do not model and analyze context explicitly. Lastly, many existing methods, such as [4, 9, 17], set strict constraints on the size of the rules or the aggregate groups to be considered, such as very small number of items/attributes allowed in a rule or only the parents and their sibling groups.

6. CONCLUSIONS AND FUTURE WORK

This paper represents the first step in an ambitious journey towards contextual outlier detection and analysis. We proposed a framework for contextual outlier detection. Our focus was to improve the interpretability of outliers. In particular, we argued that the context of an outlier should include a shared AVS, a reference group, an outlier group, and an outlier degree measure. We developed a detection algorithm leveraging the state-of-the-art data cube computation techniques.

7. REFERENCES