Identifying Outliers in Complex Categorical Data by Modelling the Feature Value Couplings

Guansong Pang[†] and **Longbing Cao**[†] and **Ling Chen**[‡]
[†]Advanced Analytics Institute, University of Technology Sydney, Australia
[‡]Quantum Computation and Intelligent Systems, University of Technology Sydney, Australia
guansong.pang@student.uts.edu.au, {longbing.cao;ling.chen}@uts.edu.au

Abstract

This paper introduces a novel unsupervised outlier detection method, namely Coupled Biased Random Walks (CBRW), for identifying outliers in categorical data with diversified frequency distributions and many noisy features. Existing pattern-based outlier detection methods are ineffective in handling such complex scenarios, as they misfit such data. CBRW estimates outlier scores of feature values by modelling feature value level couplings, which carry intrinsic data characteristics, via biased random walks to handle this complex data. The outlier scores of feature values can either measure the outlierness of an object or facilitate the existing methods as a feature weighting and selection indicator. Substantial experiments show that CBRW can not only detect outliers in complex data significantly better than the state-of-the-art methods, but also greatly improve the performance of existing methods on data sets with many noisy features.

1 Introduction

Outliers (or anomalies) are rare data objects, i.e., those objects with rare combinations of feature values, compared to the majority of objects. Detecting outliers in categorical data has wide applications in various domains, such as intrusion detection, fraud detection, and early detection of diseases, where categorical features are the only or indispensable features for describing data objects [Chandola *et al.*, 2009].

Many real-world categorical data sets own one or both of the following key characteristics: (i) there are diversified frequency distributions across different features (e.g., different mode frequencies), resulting in different semantics of the frequencies of different *patterns* (i.e., combinations of feature values); (ii) they contain a large proportion of noisy features - *features in which normal objects contain infrequent behaviours while outliers contain frequent behaviours*.

Most unsupervised outlier detection methods for categorical data (e.g., [He *et al.*, 2005; Das and Schneider, 2007; Akoglu *et al.*, 2012]) are pattern-based methods, which search for outlying/normal patterns and employ pattern frequency as a direct outlierness measure to detect outliers (i.e., patterns have the same outlierness if they have the same frequency). These methods are ineffective in handling data sets with aforementioned characteristics, as the semantic of each pattern and its frequency are different from one another and a proportion of the patterns they obtain are misleading. We call such data as *complex data* in the sense that outliers cannot be easily identified by patterns.

In a fraud detection example shown in Table 1, the feature values '*bachelor*' and '*divorced*' have the same outlierness in pattern-based methods since they have the same frequency $\frac{1}{6}$, but the frequency can indicate different outlierness in the features '*Education*' and '*Marriage*' (e.g., having different deviations from the mode frequencies $\frac{1}{2}$ and $\frac{5}{12}$, respectively). Also, by only examining the features '*Marriage*' and '*Income*', it is easy to derive some patterns (e.g., one is divorced and has low income) to successfully spot the cheater, while it is difficult to derive effective outlying/normal patterns with the presence of the noisy features '*Gender*' and '*Education*'.

Table 1: An example of fraud detection

ID	Gender	Education	Marriage	Income	Cheat?
1	male	master	divorced	low	yes
2	female	master	married	medium	no
3	male	master	single	high	no
4	male	bachelor	married	medium	no
5	female	master	divorced	high	no
6	male	PhD	married	high	no
7	male	master	single	high	no
8	female	PhD	single	medium	no
9	male	PhD	married	medium	no
10	male	bachelor	single	low	no
11	female	PhD	married	medium	no
12	male	master	single	low	no

In this paper, we introduce a new unsupervised outlier detection method, namely Coupled Biased Random Walk (CBRW), to identify outliers in those complex data. CBRW estimates the outlierness of *each feature value* by capturing both intra- and inter-feature value couplings. By intra-feature value couplings, we consider the within-feature frequency distribution to manifest the outlierness semantic of a value frequency in diversified frequency distributions. For example, the resultant outlierness of '*bachelor*' is 0.58 while that of '*divorced*' is 0.59, considering the two different frequency distributions taken by the features '*Education*' and '*Marriage*', i.e., $\{\frac{1}{2}, \frac{1}{3}, \frac{1}{6}\}$ and $\{\frac{5}{12}, \frac{5}{12}, \frac{1}{6}\}$, respectively.

By inter-feature value couplings, we score feature values based on its interactions with values of other features. Motivated by examples in diffusion networks, e.g., if someone has an overweight friend, his/her chance of becoming obese increases by 57% [Christakis and Fowler, 2007], we employ an iterative process to model outlierness propagation between feature values: a feature value has high outlierness if it has strong couplings (e.g., high conditional probabilities) with many other outlying values. This enables CBRW to distinct outlying values from noisy feature values, as noisy values are supposed to have weak couplings with outlying values. For example, compared to the noisy value '*bachelor*', although the outlying value '*low*' has lower outlierness by only considering intra-feature value couplings, it has much higher outlierness when adding inter-feature value couplings, because it has stronger couplings with the exceptional value '*divorced*'.

These two value-level couplings, which carry intrinsic data characteristics, are data-driven factors and are fed into a biased random walk model [Gómez-Gardeñes and Latora, 2008] to handle the complex data.

Note that CBRW is fundamentally different from another method that computes the value outlier scores by using the marginal probabilities of the feature values [Das and Schneider, 2007], which ignores the interactions within and between features. As a result, it fails to work on complex data sets.

Accordingly, two major contributions are made:

i. We propose a novel coupled unsupervised outlier detection method CBRW. It estimates the outlierness of each feature value by modelling intra- and inter-feature value couplings via biased random walks on an attribute graph to tackle the two aforementioned issues.

ii. The estimated outlier scores of feature values can either detect outliers directly or determine feature selection for subsequent outlier detection. To the best of our knowledge, this is the first work having such a characteristic.

Substantial experiments show that (1) our CBRW-based outlier detection method significantly outperforms the state-of-the-art methods on complex data sets; (2) without the costly pattern searching, our method runs much faster than the pattern-based methods; and (3) the CBRW-based feature selection method greatly lifts the pattern-based outlier detectors on data sets with many noisy features in terms of both accuracy and efficiency.

The rest of this paper is organised as follows. We discuss the related work in Section 2. CBRW is detailed in Section 3. The use of CBRW for outlier detection is presented in Section 4. Experiments and evaluation are provided in Section 5. We conclude this work in Section 6.

2 Related Work

Many unsupervised outlier detection methods were designed for numeric data [Breunig *et al.*, 2000; Ramaswamy *et al.*, 2000; Bay and Schwabacher, 2003]. By contrast, significantly less research has been conducted on categorical data. Existing methods on categorical data are mainly pattern-based methods, which search for infrequent/frequent patterns as outlying/normal patterns through different methods (e.g., frequent pattern mining [Ghoting *et al.*, 2004; He *et al.*, 2005; Koufakou and Georgiopoulos, 2010; Smets and Vreeken, 2011], information-theoretic measures [Akoglu *et al.*, 2012; Wu and Wang, 2013], and probability tests [Das and Schneider, 2007]) to build respective detection models. The objects with infrequent behaviours are identified as outliers. However, for a data set with many noisy features, these methods identify a large proportion of misleading patterns, leading to high false positive error. Also, the resultant patterns are derived from different feature combinations, which can have very different frequency distributions. Therefore, the semantic and importance of pattern frequency differ significantly for different patterns, and thus pattern frequency-based methods cannot appropriately capture the outlierness in data sets with diversified frequency distributions.

In addition, the pattern searching has time and space complexities that are exponential to the number of features. A heuristic search is used in [Akoglu *et al.*, 2012] to reduce the complexities from exponential to quadratic, but the search is still computationally intensive in high dimensional data.

Proper feature selection can remove noisy/irrelevant features to improve the performance of outlier detection on complex data. However, existing feature selection research focuses on regression, classification and clustering [Liu and Yu, 2005; Hesterberg *et al.*, 2008]. Very limited feature selection methods have been designed for outlier detection. The work in [Azmandian *et al.*, 2012] is an early attempt, but their proposed method is supervised and it is for numeric data.

Coupling analysis has been successfully employed to tackle complex problems in different domains, e.g., temporal outlying behaviour detection [Cao *et al.*, 2012] and similarity analysis [Wang *et al.*, 2015]. This work extends this methodology by integrating the coupling analysis via a graph-based ranking model to identify outliers in categorical data.

3 CBRW for Estimating Outlier Scores of Feature Values

The proposed method CBRW first maps the categorical data into a value-value attribute graph. The topological structure of the graph is built on inter-feature value couplings, while the node property is obtained based on intra-feature value couplings. The task of estimating the value outlierness is then transformed to solve a graph-based ranking problem (i.e., to rank the nodes). We finally build biased random walks on the graph to obtain outlier scores of feature values.

3.1 Preliminaries

Let $X = \{x_1, x_2, \cdots, x_N\}$ be a set of data objects with size N, described by a set of D categorical features $F = \{f_1, f_2, \cdots, f_D\}$. Each feature $f \in F$ has a domain $dom(f) = \{v_1, v_2, \cdots\}$, which consists of a finite set of possible feature values. The value of an object x in a feature f is denoted by $g_f(x)$. Each feature f is associated with a *categorical distribution*, where f takes on one of the possible values $v \in dom(f)$ with a frequency $p(v) = \frac{|\{x \in X | g_f(x) = v\}|}{N}$.

The domains between features are distinct, i.e., $dom(f_i) \cap dom(f_j) = \emptyset, \forall i \neq j$, and the whole set of feature values V is the union of all the feature domains: $V = \cup_{f \in F} dom(f)$.

Let $G =< V, E >$ be a directed and weighted graph, each node $v \in V$ in G represents a feature value, and entries in the **out-degree** adjacent matrix $\mathbf{A}$ of G represent weights assigned to edges. Since G is a value-value graph, the terms 'feature value' and 'node' are used interchangeably hereafter.

3.2 Node Property Based on Intra-feature Value Couplings

In the constructed graph, each node is associated with a node property, which is defined based on the value frequency and the frequency of the mode within a feature.

Definition 1. *A* mode *of a categorical distribution of a feature $f \in F$, denoted as m, is defined as a feature value $v_i \in dom(f)$ such that $p(v_i) = max(p(v_1), \cdots, p(v_K))$, where K is the number of possible values in f.*

Proposition 1. *Given any two modes m_i and m_j of features f_i and f_j, and their frequencies $p(m_i)$ and $p(m_j)$, if $p(m_i) \neq p(m_j)$, then $p(u)$ is not directly comparable to $p(v)$ in terms of their outlierness, $\forall u \in dom(f_i), v \in dom(f_j)$.*

It is assumed that outliers are rare objects compared to the majority of objects. In the same spirit, the rarity of a feature value should be evaluated against the frequencies of other values. The modes are the central tendency of the features, and their frequencies represent the majority. $p(m_i) \neq p(m_j)$ indicates that f_i and f_j have different rarity evaluation benchmarks, and thus $p(u)$ and $p(v)$ are not directly comparable.

Definition 2. *The* intra-feature outlierness *of a feature value $v \in dom(f)$ is defined by the sum of the deviation of the value frequency from the mode frequency, $dev(v)$, and the outlierness of the feature mode m, $base(m)$:*

$$\delta(v) = \frac{1}{2}\left(dev(v) + base(m)\right) \tag{1}$$

where $dev(v) = \frac{p(m)-p(v)}{p(m)}$ and $base(m) = 1 - p(m)$.

The intra-feature outlierness δ is in the range $(0, 1)$ [1]. It is defined to make values from different frequency distributions semantically comparable. In this measure, the outlierness of the feature mode is employed as a base, and the more the frequency of a feature value deviates from the mode frequency, the more outlying the value is. Any function that guarantees a monotonic decreasing relation between $base(m)$ and $p(m)$ can be used to compute the outlierness of m. Our empirical results show that $base(m) = 1 - p(m)$ performs more stably than other functions, so we use this specification.

3.3 Graph Topological Structure Based on Inter-feature Value Couplings

We extract information from the inter-feature value couplings to examine whether the values of a feature are coupled with the outlying behaviours of the rest of features.

Definition 3. *The entry $\mathbf{A}(u, v)$ is a weight assigned to the edge from node u to node v and is defined as:*

$$\mathbf{A}(u, v) = p(u|v) = \frac{p(u, v)}{p(v)} \tag{2}$$

where $p(u, v)$ denotes the co-occurrence frequency of the values u and v, $\forall u, v \in V$.

The entry $\mathbf{A}(u, v)$ can be interpreted as that outlierness is propagated from u to v at the rate of $p(u|v)$. In other words, if u, v has a strong coupling and u has high outlierness, the outlierness propagating from u to v would be high; and v has high outlierness if and only if there are many nodes having a similar relation as u to v. In this sense, the outlierness of v is also dependent on the behaviours of its co-occurred values. This makes our method less sensitive to noisy features. Note that if u and v are values of the same feature, we have $\mathbf{A}(u, v) = 0$ due to $p(u, v) = 0$. Thus, $\mathbf{A}$ can be regarded as a matrix of *inter-feature outlierness*.

The adjacent matrix $\mathbf{A}$ is built on conditional probabilities between every two nodes. This is different from the conventional adjacent matrix construction that fills the matrix based on similarities between the nodes. Our way is because the conditional probabilities are simple and they fully capture the desired co-occurrence behaviours, while the similarities between two values in different features are not well defined.

3.4 Biased Random Walks on the Attribute Graph

In building *unbiased* random walks (URWs), after obtaining $\mathbf{A}$, we can then obtain *transition matrix* $\mathbf{W}$ as:

$$\mathbf{W} = \mathbf{A}\mathbf{D}^{-1} \tag{3}$$

where $\mathbf{D}$ is the diagonal matrix of $\mathbf{A}$ with its u-th diagonal entry $d(u) = \sum_{v \in V} \mathbf{A}(u, v)$. The entry $\mathbf{W}(u, v) = \frac{\mathbf{A}(u,v)}{d(u)}$, which represents the probability of the transition from node u to node v, and it satisfies that $\sum_{v \in V} \mathbf{W}(u, v) = 1$. This transition matrix neglects the intra-feature outlierness, i.e., the node property.

Here we build *biased* random walks (BRWs) to introduce a bias into the random walking process, and define the entry of the transition matrix as follows:

$$\mathbf{W}_b(u, v) = \frac{\delta(v)\mathbf{A}(u, v)}{\sum_{v \in V} \delta(v)\mathbf{A}(u, v)} \tag{4}$$

$\mathbf{W}_b(u, v)$ indicates that the transition from node u to node v has a probability proportional to $\delta(v)\mathbf{A}(u, v)$. Therefore, every random move is biased by the values of δ associated with the nodes.

Let the column vector $\boldsymbol{\pi}_t \in \mathbb{R}^{|V|}$ denote the *probability distribution* of the biased random walk at time step t, i.e., the probability of a random walker visiting any given node at the t-th step. Then we have:

$$\boldsymbol{\pi}_{t+1} = \mathbf{W}_b^\intercal \boldsymbol{\pi}_t \tag{5}$$

where $[\cdot]^\intercal$ is the matrix transpose operation, that is, $\boldsymbol{\pi}_{t+1}(v) = \sum_{u \in V} \mathbf{W}_b(u, v)\boldsymbol{\pi}_t(u)$.

Corollary 1. *If G is irreducible and aperiodic, $\boldsymbol{\pi}$ converges to a unique stationary probability $\boldsymbol{\pi}^*$ s.t. $\boldsymbol{\pi}^* = \mathbf{W}_b^\intercal \boldsymbol{\pi}^*$.*

Proof. It is easy to see that BRWs using $\mathbf{W}_b$ is equivalent to URWs on a graph G_e with an adjacent matrix $\mathbf{A}_e$, where the entry $\mathbf{A}_e(u, v) = \delta(u)\mathbf{A}(u, v)\delta(v)$, as the transition matrix $\mathbf{W}_e$ of G_e satisfies: $\mathbf{W}_e \equiv \mathbf{W}_b$.

[1] We neglect features which have $p(m) = 1$, as those features contain no information for outlier detection.

Since δ is always positive, $\mathbf{A}_e$ and $\mathbf{A}$ have the same irreducibility and aperiodicity [Aldous and Fill, 2002]. If G is irreducible and aperiodic, so does G_e. Based on the Perron–Frobenius Theorem [Aldous and Fill, 2002], we have $\boldsymbol{\pi}^* = \mathbf{W}_e^\intercal \boldsymbol{\pi}^* = \mathbf{W}_b^\intercal \boldsymbol{\pi}^*$. ■

This states that the stationary probabilities of nodes are independent of initialisations of $\boldsymbol{\pi}$, and they are proportional to the in-degree weights of the nodes. Motivated by this, we define the outlier score of a feature value as follows.

Definition 4. *The outlier score of node v is defined by its stationary probability:*

$$value\_score(v) = \boldsymbol{\pi}^*(v) \tag{6}$$

where $0 < \boldsymbol{\pi}^*(v) < 1$ and $\sum_{v \in V} \boldsymbol{\pi}^*(v) = 1$.

The value v has a large outlier score if and only if it demonstrates outlying behaviour within the feature as well as co-occurs with many outlying values, since $\boldsymbol{\pi}^*(v)$ is proportional to $\mathbf{W}_b(u, v)$, which is determined by $\delta(v)$ and $\mathbf{A}(u, v)$.

3.5 The Algorithm of CBRW

Algorithm 1 presents the procedures of the coupled biased random walk model (CBRW). Steps (1-8) are performed to obtain the information of the intra- and inter-feature value couplings. The matrix $\mathbf{W}_b$ is then generated based on Equations (2) and (4).

Algorithm 1 *Estimate Value Outlierness (X, α)*

Input: X - data objects, α - damping factor
Output: $\boldsymbol{\pi}^*$ - the stationary probability distribution

1: **for** $i = 1$ to D **do**
2: Compute $p(v)$ for each $v \in dom(f_i)$
3: Find the mode of f_i
4: Compute $\delta(v)$
5: **for** $j = i + 1$ to D **do**
6: Compute $p(u, v)$, $\forall u \in dom(f_j)$
7: **end for**
8: **end for**
9: Generate the matrix $\mathbf{W}_b$
10: Initialise $\boldsymbol{\pi}^*$ as a uniform distribution
11: **repeat**
12: $\boldsymbol{\pi}^* \leftarrow (1 - \alpha)\frac{1}{|V|}\mathbf{1} + \alpha \mathbf{W}_b^\intercal \boldsymbol{\pi}^*$
13: **until** Convergence, i.e., $||\boldsymbol{\pi}_t^* - \boldsymbol{\pi}_{t-1}^*||_\infty \leq 0.001$ or reach the maximum iteration $I_{max} = 100$
14: **return** $\boldsymbol{\pi}^*$

Note that our constructed graph could be reducible (e.g., if there are isolated groups of nodes) or periodic (e.g., G is a bipartite graph if a data set has two features only). In Step (12), following [Page *et al.*, 1998], a damping factor α, is introduced into Equation (5) to guarantee the convergence:

$$\boldsymbol{\pi}_{t+1} = (1 - \alpha)\frac{1}{|V|}\mathbf{1} + \alpha \mathbf{W}_b^\intercal \boldsymbol{\pi}_t \tag{7}$$

We employ $\alpha = 0.95$ in our experiments, but our empirical results show that CBRW performs stably with the varying of α, e.g., $\alpha \in [0.85, 0.99]$.

CBRW requires only one scanning over the data objects to obtain the value couplings information in Steps (1-8). The random walks in Steps (11-13) has $O(|E|I_{max})$. The data size N is often far larger than $|E|I_{max}$, so the runtime is determined by N, which is linear to the data size. Theoretically, our method has $O(D^2)$, as two loops are required in order to obtain the value co-occurrence information. However, the computation within the inner loop (i.e., Step (6)) is just a simple counting, leading to a nearly linear time complexity to the number of features in practice.

4 Outlier Detection using CBRW

Two basic applications of outlier scores of feature values are the vertical and horizontal summation.

4.1 Feature Weighting and Selection for Outlier Detection

Vertically, the sum of the outlier scores of the values in a feature can be utilised to weight and select features.

In outlier detection, relevant features are the features where the outliers demonstrate outlying behaviours and are distinguishable from normal objects. Thus, the feature relevance can be measured by the sum of outlierness carried by each value of the feature.

Definition 5. *The relevance of a feature f is defined as:*

$$rel(f) = \sum_{v \in dom(f)} value\_score(v) \tag{8}$$

Large $rel(\cdot)$ indicates high relevance of the feature to outlier detection. Top-ranked features are the most relevant features, while the bottom-ranked are noisy/irrelevant features.

Our CBRW-based feature selection method (denoted as `CBRW_FS`) selects top-ranked features for each data set. Outlier detectors can then work on the newly obtained data sets with the selected features for enhancing their robustness to noisy features and/or reducing their computation time on high dimensional data. These relevance weights can also be embedded in the outlier scoring function of an outlier detector as a feature weighting.

4.2 Direct Outlier Detection

Horizontally, the outlier scores of feature values can measure the outlierness of an object as follows.

Definition 6. *The outlier score of an object x is defined as:*

$$object\_score(x) = \sum_{f \in F} w_f \times value\_score(g_f(x)) \tag{9}$$

where $w_f = \frac{rel(f)}{\sum_{f \in F} rel(f)}$ is a feature weighting component.

The outlier score of an object is the weighted sum of the outlier scores of the values contained by the object, with a relevance weighting factor to highlight the importance of $value\_score(\cdot)$ in highly relevant features.

Our CBRW-based outlier detection method (denoted as `CBRW_OD`) employs Equation (9) to rank the data objects. Outliers are data objects having large outlier scores.

The component $value\_score(\cdot)$ has taken account of the diversified frequency distributions and noisy feature issues. The feature weighting in Equation (9) can further enhance the robustness of our outlier detection method to noisy features.

5 Experiments

5.1 Outlier Detectors and Its Parameter Settings

Our method CBRW was evaluated against three state-of-the-art outlier detectors: the frequent pattern-based FPOF [He *et al.*, 2005], the information-theoretic-based CompreX (denoted as COMP) [Akoglu *et al.*, 2012], and the probability test-based MarP [Das and Schneider, 2007].

CBRW used $\alpha = 0.95$ by default. Following [He *et al.*, 2005], FPOF was used with the minimum *support* threshold $min\_sup = 0.1$ and the maximum pattern length $l = 5$. Both COMP and MarP are parameter-free.

CBRW, FPOF and MarP were implemented in JAVA in WEKA [Hall *et al.*, 2009]. COMP was obtained from the authors of [Akoglu *et al.*, 2012] in MATLAB. All the experiments were performed at a node in a 3.4GHz Phoenix Cluster with 32GB memory.

5.2 Performance Evaluation Method

All the outlier detectors produce a ranking based on their outlier scores, i.e., top ranked objects are the most likely outliers. The area under ROC curve (AUC) was derived based on the ranking [Hand and Till, 2001]. Higher AUC indicates better accuracy. We compare the efficiency in our scale-up test.

A commonly used evaluation method for unsupervised learning was taken here, namely, detectors were trained and evaluated on the same data set, but it was assumed that the ground truth is unavailable in the training. The ground truth was only involved in computing the AUC in the evaluation.

5.3 Data Sets

Eleven publicly available real-world data sets were used, which cover diverse domains, e.g., intrusion detection, text classification and image object recognition, as shown in Table 2. *Probe* and *U2R* were derived from KDDCUP99 data sets using probe and user-to-root attacks as outliers against the normal class, respectively. Other data sets were transformed from extremely imbalanced data, where the rare classes were treated as outliers versus the rest of classes as normal class [Lazarevic and Kumar, 2005; Chen *et al.*, 2009]. The percentage of outliers in each data set ranges from 0.1% to 6.2%.

Data factor refers to underlying data characteristics of data sets, which are associated with the detection performance of outlier detectors. Two key data factors are presented below, and their quantisation results are reported in Table 2.

- Feature noise level κ_{noise}. We computed the AUC using MarP for each individual feature. A feature is regarded as a noisy feature if the AUC is less than 0.5. We report the percentage of noisy features.
- Variation among the frequencies of modes κ_{mode}. The modes of all the D features were sorted based on their frequencies in descending order $\{m_{k_1}, \cdots, m_{k_D}\}$ where $1 \leq k_1 < k_D \leq D$, and an average variation extent is obtained as: $\frac{2}{D(D-1)} \sum_{k_i<k_j} \frac{p(m_{k_i})}{p(m_{k_j})} \in [1, \infty)$.

5.4 Evaluation Results

The AUC results of CBRW_OD, FPOF, COMP and MarP on 11 data sets with different κ_{noise} and κ_{mode} are presented in Table 2. The p-value results are based on paired two-tailed t-test using the null hypothesis that the AUC results of CBRW and another detector come from distributions with equal means.

Table 2: AUC Results of Four Detectors on 11 Data Sets. SF, CT, R10 and Link are short for *Solar Flare*, *CoverType*, *Reuters10* and *Linkage*, respectively. CBRW is short for CBRW_OD. The horizontal line in the middle is a rough separation between complex and simple data. FPOF runs out-of-memory in high dimensional data sets *aPascal* and *Reuters10*. The best performance for each data set is boldfaced.

Basic Data Info.			**Data Factors**		**Outlier Detectors**			
Name	N	D	κ_{noise}	κ_{mode}	CBRW	FPOF	COMP	MarP
aPascal	12,695	64	81%	1.19	**0.82**	NA	0.66	0.62
Census	299,285	33	58%	1.65	**0.67**	0.61	0.64	0.59
CelebA	202,599	39	49%	1.26	**0.85**	0.74	0.76	0.74
CMC	1,473	8	37%	1.58	**0.63**	0.56	0.57	0.54
Chess	28,056	6	33%	2.24	**0.79**	0.62	0.64	0.64
SF	1,066	12	8%	1.55	**0.88**	0.86	0.85	0.84
Probe	64,759	7	0%	1.31	**0.99**	**0.99**	0.98	0.98
Link	5,749,132	5	0%	1.39	1.00	1.00	1.00	1.00
R10	12,897	100	23%	1.03	0.99	NA	0.99	0.99
CT	581,012	44	34%	1.10	0.97	**0.98**	**0.98**	**0.98**
U2R	60,821	7	14%	1.27	0.97	0.92	**0.99**	0.88
					p-value	0.027	0.034	0.007

Detection Performance Summary

CBRW_OD achieves the best detection performance on seven data sets, and performs equally well with all other detectors on two data sets, with two close to the best (having the difference in AUC no more than 0.02). The significance test results show that CBRW_OD outperforms its three contenders significantly at the 95% confidence level.

Handling Data Sets with High κ_{noise}

CBRW_OD performs substantially better than the other three detectors in all the four data sets with high κ_{noise} (e.g., $\kappa_{noise} > 35\%$) (i.e., *aPascal*, *Census*, *CelebA* and *CMC*). On average, it obtains more than 12%, 13% and 19% AUC improvement over FPOF, COMP and MarP, respectively.

We also evaluated our method in terms of handling noisy features by using it to select features (e.g., removing noisy/irrelevant features) for subsequent outlier detection.

Figure 1 shows that the AUC results of the four detectors with or without using our feature selection method (CRBW_FS) on *Census* and *aPascal*, which have the largest percentage of noisy features. All the four enhanced detectors can obtain substantial AUC improvement by working on data sets with selected top-ranked features over a wide range of selection options.

To observe the effect of using CRBW_FS on AUC performance and runtime, we examined the results over a range of feature selection options indicated by the inclusive areas of vertical pink lines in Figure 1 and report the average results in Table 3.

Table 3 shows that, on average, FPOF, COMP and MarP can obtain 5% to 36% AUC improvement. Since CBRW_OD

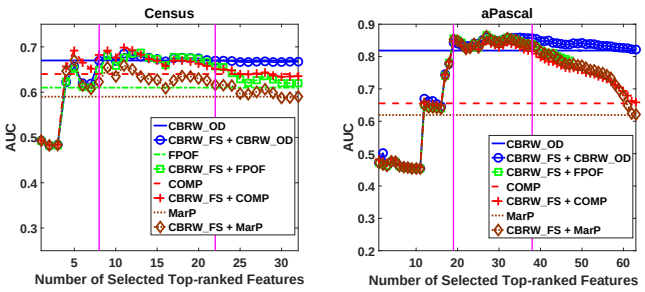

Figure 1: AUC Results of the Four Detectors on *Census* and *aPascal* Using Our Feature Selection method CRBW_FS. The performance on the original data is used as baseline; Note that FPOF runs out of memory in *aPascal*.

has the ability to handle noisy features without using feature selection, its AUC improvement is much smaller than the other three detectors. Also, by working on data sets with reduced number of features, some detectors might obtain substantial efficiency improvement, e.g., on average, FPOF can run four orders of magnitude faster on *Census* with the top-ranked features than that working on the full feature set.

Table 3: Improvement by Using CRBW_FS in the Four Detectors on *Census* and *aPascal*. CBRW is short for CBRW_OD.

	AUC Improvement				Speedup Ratio			
	CBRW	FPOF	COMP	MarP	CBRW	FPOF	COMP	MarP
Census	1%	10%	5%	7%	2.23	1485.73	1.24	2.35
aPascal	4%	NA	27%	36%	2.74	NA	6.5	1.35

Handling Data Sets with High κ_{mode}

CBRW_OD outperforms FPOF, COMP and MarP substantially on all the four data sets with high κ_{mode} (e.g., $\kappa_{mode} > 1.50$) (i.e., *Census*, *CMC*, *Chess* and *Solar Flare*). On average, it obtains more than 13%, 11% and 14% AUC improvement over FPOF, COMP and MarP, respectively.

Handling Simple Data Sets with Low κ_{noise} and κ_{mode}

All the four detectors perform very well on data sets with low κ_{noise} and κ_{mode}. This is particularly true for data sets with extremely low κ_{noise} or κ_{mode}. For example, all the four detectors, including the most simple detector MarP, obtain the AUC of (or nearly) one on *Linkage* and *Probe* with $\kappa_{noise} = 0$, and *Reuters10* and *CoverType* with $\kappa_{mode} \leq 1.10$.

Scalability Test

The scale-up test results are presented in Figure 2. The results reported in the left panel show that all the four detectors have runtime linear to data size. The runtime of COMP increases by a factor of more than 3,000 when the data size increases by a factor of 256; while that of CBRW_OD increases by less than 60. Therefore, though CBRW_OD and COMP were implemented in different programming languages, the difference in *runtime ratio* indicates that CBRW_OD runs much faster than COMP by a factor of more than 50.

The results reported in the right panel in Figure 2 show that CBRW_OD runs more than five orders of magnitude faster than

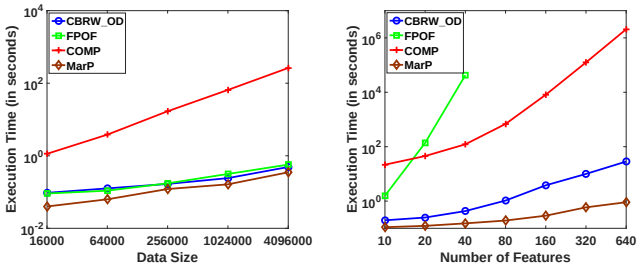

Figure 2: Scale-up Test Results of the Four Detectors w.r.t. Data Size and Dimensionality. Note that FPOF runs out-of-memory when the number of features reaches 80.

FPOF, and it runs slower than MarP by a factor of more than 30. As indicated by runtime ratio, CBRW_OD runs much faster than COMP by a factor of more than 500.

5.5 Discussions

Below, we briefly discuss the impact of feature weighting and other data factors.

- **Feature weighting**. The outlier scoring function in Equation (9) includes a feature weighting component. The empirical results show that the unweighted version performs slightly less effectively than the weighted version on the four data sets with high feature noise level, but it still substantially outperforms the three contenders on these four data sets, and it has the same performance as the weighted version on other data sets. Those results are omitted due to the space limit.
- **Other key data factors**. There are some other key data factors, e.g., the minimum length of outlying patterns. That is, some outliers are detectable only by looking at combinations of no less than k features. In data sets where the pattern length factor is a dominant factor, e.g., *U2R*, our outlier detector performs less effective than the pattern-based methods that search for patterns of all possible lengths (e.g., COMP).

6 Conclusions

This paper introduces a new unsupervised outlier detection method CBRW for detecting outliers in complex categorical data. Compared to the pattern-based methods, CBRW is data-driven, which learns from low-level intra- and inter-feature value couplings to estimate outlier scores of feature values. Substantial experiments show that our CBRW-based outlier detector (CBRW_OD) can significantly outperform other detectors including FPOF, COMP and MarP on complex data. Further, our CBRW-based feature selection method (CRBW_FS) can greatly improve the performance of existing detectors on data sets with many noisy features. CBRW_OD runs more than two to five orders of magnitude faster than COMP and FPOF. We are further enhancing the efficiency of CBRW, and extending its applicability, e.g., by integrating the pattern length factor.

Acknowledgements

We would like to thank anonymous reviewers for their helpful and constructive comments. This work is partially supported by the ARC Discovery Project under Grant NO.DP130102691 and No.DP140100545.

References

[Akoglu *et al.*, 2012] Leman Akoglu, Hanghang Tong, Jilles Vreeken, and Christos Faloutsos. Fast and reliable anomaly detection in categorical data. In *CIKM*, pages 415–424. ACM, 2012.

[Aldous and Fill, 2002] David Aldous and Jim Fill. *Reversible Markov chains and random walks on graphs*. Berkeley, 2002.

[Azmandian *et al.*, 2012] Fatemeh Azmandian, Ayse Yilmazer, Jennifer G Dy, Javed Aslam, David R Kaeli, et al. GPU-accelerated feature selection for outlier detection using the local kernel density ratio. In *ICDM*, pages 51–60. IEEE, 2012.

[Bay and Schwabacher, 2003] Stephen D Bay and Mark Schwabacher. Mining distance-based outliers in near linear time with randomization and a simple pruning rule. In *SIGKDD*, pages 29–38. ACM, 2003.

[Breunig *et al.*, 2000] Markus M Breunig, Hans-Peter Kriegel, Raymond T Ng, and Jörg Sander. LOF: Identifying density-based local outliers. *ACM SIGMOD Record*, 29(2):93–104, 2000.

[Cao *et al.*, 2012] Longbing Cao, Yuming Ou, and Philip S Yu. Coupled behavior analysis with applications. *IEEE Transactions on Knowledge and Data Engineering*, 24(8):1378–1392, 2012.

[Chandola *et al.*, 2009] Varun Chandola, Arindam Banerjee, and Vipin Kumar. Anomaly detection: A survey. *ACM Computing Surveys*, 41(3):15, 2009.

[Chen *et al.*, 2009] Yixin Chen, Xin Dang, Hanxiang Peng, and Henry L Bart. Outlier detection with the kernelized spatial depth function. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 31(2):288–305, 2009.

[Christakis and Fowler, 2007] Nicholas A Christakis and James H Fowler. The spread of obesity in a large social network over 32 years. *New England Journal of Medicine*, 357(4):370–379, 2007.

[Das and Schneider, 2007] Kaustav Das and Jeff Schneider. Detecting anomalous records in categorical datasets. In *SIGKDD*, pages 220–229. ACM, 2007.

[Ghoting *et al.*, 2004] Amol Ghoting, Matthew Eric Otey, and Srinivasan Parthasarathy. LOADED: Link-based outlier and anomaly detection in evolving data sets. In *ICDM*, pages 387–390. IEEE, 2004.

[Gómez-Gardeñes and Latora, 2008] Jesús Gómez-Gardeñes and Vito Latora. Entropy rate of diffusion processes on complex networks. *Physical Review E*, 78(6):065102, 2008.

[Hall *et al.*, 2009] Mark Hall, Eibe Frank, Geoffrey Holmes, Bernhard Pfahringer, Peter Reutemann, and Ian H Witten. The WEKA data mining software: An update. *ACM SIGKDD Explorations Newsletter*, 11(1):10–18, 2009.

[Hand and Till, 2001] David J Hand and Robert J Till. A simple generalisation of the area under the ROC curve for multiple class classification problems. *Machine Learning*, 45(2):171–186, 2001.

[He *et al.*, 2005] Zengyou He, Xiaofei Xu, Zhexue Joshua Huang, and Shengchun Deng. FP-outlier: Frequent pattern based outlier detection. *Computer Science and Information Systems*, 2(1):103–118, 2005.

[Hesterberg *et al.*, 2008] Tim Hesterberg, Nam Hee Choi, Lukas Meier, Chris Fraley, et al. Least angle and $\ell 1$ penalized regression: A review. *Statistics Surveys*, 2:61–93, 2008.

[Koufakou and Georgiopoulos, 2010] Anna Koufakou and Michael Georgiopoulos. A fast outlier detection strategy for distributed high-dimensional data sets with mixed attributes. *Data Mining and Knowledge Discovery*, 20(2):259–289, 2010.

[Lazarevic and Kumar, 2005] Aleksandar Lazarevic and Vipin Kumar. Feature bagging for outlier detection. In *SIGKDD*, pages 157–166. ACM, 2005.

[Liu and Yu, 2005] Huan Liu and Lei Yu. Toward integrating feature selection algorithms for classification and clustering. *IEEE Transactions on Knowledge and Data Engineering*, 17(4):491–502, 2005.

[Page *et al.*, 1998] Lawrence Page, Sergey Brin, Rajeev Motwani, and Terry Winograd. The PageRank citation ranking: Bringing order to the web. In *WWW Conference*, pages 161–172, 1998.

[Ramaswamy *et al.*, 2000] Sridhar Ramaswamy, Rajeev Rastogi, and Kyuseok Shim. Efficient algorithms for mining outliers from large data sets. *ACM SIGMOD Record*, 29(2):427–438, 2000.

[Smets and Vreeken, 2011] Koen Smets and Jilles Vreeken. The odd one out: Identifying and characterising anomalies. In *SDM*, pages 109–148. SIAM, 2011.

[Wang *et al.*, 2015] Can Wang, Xiangjun Dong, Fei Zhou, Longbing Cao, and Chi-Hung Chi. Coupled attribute similarity learning on categorical data. *IEEE Transactions on Neural Networks and Learning Systems*, 26(4):781–797, 2015.

[Wu and Wang, 2013] Shu Wu and Shengrui Wang. Information-theoretic outlier detection for large-scale categorical data. *IEEE Transactions on Knowledge and Data Engineering*, 25(3):589–602, 2013.