Selective Value Coupling Learning for Detecting Outliers in High-Dimensional Categorical Data

Guansong Pang
University of Technology Sydney
Sydney, Australia
pangguansong@gmail.com

Hongzuo Xu
National University of Defense Technology
Changsha, China
leogarcia@126.com

Longbing Cao
University of Technology Sydney
Sydney, Australia
longbing.cao@uts.edu.au

Wentao Zhao
National University of Defense Technology
Changsha, China
wtzhao@nudt.edu.cn

ABSTRACT

This paper introduces a novel framework, namely SelectVC and its instance POP, for learning *selective value couplings* (i.e., interactions between the full value set and a set of outlying values) to identify outliers in high-dimensional categorical data. Existing outlier detection methods work on a full data space or feature subspaces that are identified independently from subsequent outlier scoring. As a result, they are significantly challenged by overwhelming irrelevant features in high-dimensional data due to the noise brought by the irrelevant features and its huge search space. In contrast, SelectVC works on a clean and condensed data space spanned by selective value couplings by jointly optimizing *outlying value selection* and *value outlierness scoring*. Its instance POP defines a value outlierness scoring function by modeling a partial outlierness propagation process to capture the selective value couplings. POP further defines a top-k outlying value selection method to ensure its scalability to the huge search space. We show that POP (i) significantly outperforms five state-of-the-art full space- or subspace-based outlier detectors and their combinations with three feature selection methods on 12 real-world high-dimensional data sets with different levels of irrelevant features; and (ii) obtains good scalability, stable performance w.r.t. k, and fast convergence rate.

KEYWORDS

Outlier Detection; High-Dimensional Data; Categorical Data; Feature Selection; Coupling Learning

1 INTRODUCTION

Outliers are rare objects, compared to the majority of normal objects. Detecting outliers plays a vital role in numerous applications, such as detecting network intrusion attacks, credit card frauds, rare diseases and social events etc. However, identifying outliers is a challenging task, in particular for complex data.

This work focuses on the problem of detecting outliers in high-dimensional categorical data. Such data poses the following two major challenges: (i) It often contains a complex mixture of relevant and irrelevant features. The irrelevant features are 'noise' to outlier detection, since outliers are masked as normal objects by these features. Moreover, the sophisticated *couplings* [9] (e.g., different types and hierarchies of interactions) within irrelevant features and between relevant and irrelevant features bring about substantially more 'noise' that impedes the separability of outliers from normal objects. (ii) It also presents a huge search space, i.e., 2^D where D is the number of features, resulting in great difficulty in exploring the mixed couplings across the features.

Most outlier detection methods (e.g., [3, 4, 13, 22]) for categorical data are subspace-based methods. These methods consist of two successive independent modules - pattern/subspace discovery and outlier scoring. In general, they first identify a set of patterns (i.e., value combinations) or subspaces, and then aggregate the *outlierness* (i.e. outlier score) in the subspaces to obtain object outlierness. Such modular design enables the application of state-of-the-art subspace/pattern discovery methods into outlier detection. However, their pattern/subspace search works separately from outlier scoring, and thus may be misled by irrelevant features and produce faulty patterns/subspaces [20]. Also, such search is very costly on high-dimensional data due to its huge search space.

In addition, there have been some full space-based methods (e.g., [5, 11, 16, 23]) using new outlier scoring functions (e.g., angles between distance vectors [16, 23]) to overcome the effect of irrelevant features. However, they work on the full feature set and become ineffective when outliers are only detectable in small feature subsets.

Feature selection has been an enabling technique for learning methods to handle high-dimensional data. Therefore, the above subspace- or full space-based methods may be empowered by feature selection to deal with irrelevant features. However, although feature selection for classification and clustering tasks has been intensively studied, only limited work [20, 21] has been done on feature selection for outlier detection. Moreover, the methods in [20, 21] perform feature selection independently from subsequent outlier scoring and may retain features that are irrelevant to the outlier scoring functions.

CIKM'17 , November 6–10, 2017, Singapore, Singapore
© 2017 Association for Computing Machinery.
ACM ISBN 978-1-4503-4918-5/17/11…$15.00
https://doi.org/10.1145/3132847.3132994

The above analysis suggests that how to effectively and efficiently identify and model on a clean and condensed space from the original data space is the key to detecting high-dimensional outliers. Accordingly, this paper proposes a novel high-dimensional outlier detection framework for categorical data by modeling *Selective Value Couplings* (the SelectVC framework for short), i.e., selective feature value interactions that are positively related to outlier detection. As shown in Figure 1, given an initial value outlierness vector which contains outlier scores of all feature values, SelectVC first defines a *value subset evaluation function* ψ to select a subset of values that are the most likely outlying values. *Outlying values* are infrequent values which are mainly contained by outliers. SelectVC then defines an *outlier scoring function* ϕ to re-compute an outlier score of every single value based on the couplings between this single value and the selected outlying value set. The scoring function ϕ models only selective value couplings in a condensed space in the sense that it focuses on the couplings of the single value with the outlying value set rather than the full value set. These two steps are iteratively performed until the value outlierness vector converges.

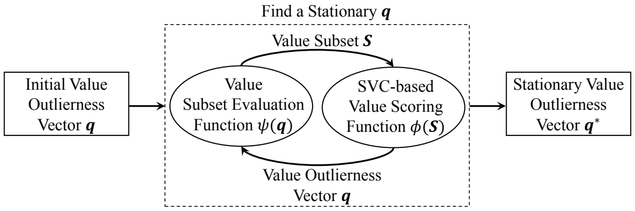
Find a Stationary $\boldsymbol{q}$
Value Subset $\boldsymbol{S}$
Initial Value Outlierness Vector $\boldsymbol{q}$
Value Subset Evaluation Function $\psi(\boldsymbol{q})$
SVC-based Value Scoring Function $\phi(\boldsymbol{S})$
Stationary Value Outlierness Vector $\boldsymbol{q}^*$
Value Outlierness Vector $\boldsymbol{q}$

Figure 1: The SelectVC Framework for Estimating Value Outlierness Based on Selective Value Couplings. The outlierness of data objects can then be obtained using value outlierness. SVC is short for Selective Value Couplings.

Outliers often demonstrate multiple outlying *behaviors* ('behaviors' and 'values' are used interchangeably hereafter) in high-dimensional data, i.e., outlying behaviors are often concurrent. Moreover, outlying behaviors have very low individual frequency. This results in strong mutual couplings between outlying behaviors. On the other hand, although outlying behaviors also co-occur with *non-outlying behaviors* (including *normal behaviors* and *noisy behaviors* - frequent and infrequent values which are mainly contained by normal objects, respectively), non-outlying behaviors are distributed very differently from outlying behaviors since they are manifested by respective normal objects and outliers. This results in weak couplings between non-outlying behaviors and outlying behaviors. The strength of couplings between outlying behaviors is therefore *contrasting* to that between non-outlying behaviors and outlying behaviors. SelectVC essentially models such contrasting couplings to iteratively assign substantially larger outlierness to outlying values than normal/noisy values. The efficiency of SelectVC is mainly determined by the value selection function (ψ).

We further instantiate the SelectVC framework to a Partial Outlierness Propagation-based method, called POP. POP specifies the scoring function ϕ by simulating partial outlierness propagation from the value subset to the full value set. POP further specifies ψ by a top-k outlying value selection function to simplify the value selection and ensure its scalability to very high-dimensional data.

This work makes the following two major contributions:

- The proposed SelectVC framework for outlier detection is novel for high-dimensional categorical data. Different from existing approaches that primarily work on the original full space and/or feature subsets identified independently from outlier scoring, SelectVC works on a clean and condensed data space composed by the couplings between the outlying value set and the full value set, by jointly optimizing outlying value selection and value outlierness scoring. This enables SelectVC to have a more reliable outlierness estimation on data with overwhelming irrelevant features.
- The performance of SelectVC is verified by its instance POP. POP models the contrasting couplings between outlying-to-outlying values and normal/noisy-to-outlying values by partial outlierness propagation. Our theoretical analysis shows that such outlierness propagation biases towards outlying behaviors, which assists POP to assign larger outlierness to outlying behaviors than non-outlying behaviors.

Extensive experiments show that POP (i) significantly outperforms five state-of-the-art full space- or subspace-based outlier detectors and their combinations with three feature selection methods (5%-39% AUC improvement) on 12 real-world high-dimensional data sets with different levels of irrelevant features; (ii) obtains good scalability w.r.t. data size and dimensionality; (iii) performs stably w.r.t. its only parameter k; and (iv) obtains fast convergence rate.

In the rest of this paper, we discuss the related work in Section 2. SelectVC is detailed in Section 3. POP is introduced in Section 4, followed by a theoretical analysis in Section 5. Empirical results are provided in Section 6. We conclude this work in Section 7.

2 RELATED WORK

Existing high-dimensional outlier detection methods can be generally categorized as deterministic and non-deterministic subspace-based methods, full space-based methods and feature selection-based methods.

Deterministic Subspace-based Methods. These subspace methods includes local pattern-based methods [2, 4, 13], feature partition-based methods [3] and statistical dependence-based methods [15]. They are deterministic in the sense that they produce exactly the same subspaces/patterns that satisfy a given criterion. They normally first search occurrence frequency/local density, minimum description length or statistical dependence tests-based outlying subspaces/patterns, and then computes outlier scores in subspaces to avoid the inclusion of irrelevant features. However, their subspace/pattern search has prohibitive computational time and/or storage requirement in high-dimensional data. Also, the presence of irrelevant features may mislead the search to produce irrelevant subspaces/patterns, leading to false positive errors [20].

Non-deterministic Subspace-based Methods. In contrast to deterministic methods, non-deterministic methods [17, 19, 22, 24] work on randomly generated subspaces. These methods generally have substantially better efficiency than deterministic methods, since they do not require the costly subspace search and their random subspace generation is very fast. However, the random subspace generation may include many irrelevant features into subspaces while omit relevant features in high-dimensional data, where irrelevant features dominant over relevant features.

Full Space-based Methods. Traditional outlier detection methods like LOF, kNN and their numerous variants [8] rely on pairwise distances on the full data space to define outliers and they fail in high-dimensional data due to the curse of dimensionality [26]. Some methods [5, 11, 16, 23] attempt to address this problem by designing new outlier definitions for high-dimensional data. Although they sucessfully avoid to directly use pairwise distance in outlier scoreing, their premises are dependent on the proximity concept in the original full space, and thus they are still biased by irrelevant features [26]. Also, these methods often require an input for the neighborhood size, which is heavily dependent on data size and data distribution and is difficult to be tuned as class labels are unavailable [22]. Different from the above methods, the methods reported in [10, 20] avoid the distance computation by using value interactions to estimate the outlierness of values/value pairs. These two methods are good at capturing complex value interactions while obtain quite good efficiency, but its performance can be considerably biased by irrelevant features, since they work on original full space and the overwhelming irrelevant features downgrade their outlierness estimation.

Feature Selection-based Methods. Feature selection has shown effective in enabling clustering and (imbalanced) classification on high-dimensional data [6, 18], but there exists limited work on outlier detection. Building on coupling learning of outliers, the methods in [20, 21] attempt to estimate the outlierness of values by learning the underlying value interactions, and use the value outlier scores to infer the relevance of features to outlier detection. However, these methods work independently from subsequent outlier detection methods and use the full value couplings to compute the outlier scores, and thus the feature selection may be biased by irrelevant value couplings, resulting in suboptimal feature subsets.

3 THE SelectVC FRAMEWORK

SelectVC jointly optimizes value selection and value outlierness scoring, which is described as follows. Let $\mathcal{X} = \{\boldsymbol{x}_1, \boldsymbol{x}_2, \cdots, \boldsymbol{x}_N\}$ be a set of data objects with size N, described by D features $\mathcal{F} = \{f_1, f_2, \cdots, f_D\}$; $dom(f) = \{v_1, v_2, \cdots\}$ be the domain of a feature f and $\mathcal{V}$ be the whole set of feature values in $\mathcal{F}$: $\mathcal{V} = \cup_{f \in \mathcal{F}} dom(f)$, where $dom(f) \cap dom(f') = \emptyset, \forall f \neq f'$. As shown in Figure 1, given an initial value outlierness vector $\mathbf{q} \in \mathbb{R}^{|\mathcal{V}|}$, SelectVC first defines a value selection function $\psi(\mathbf{q})$ to select a set of outlying values, $\mathcal{S} \subset \mathcal{V}$. SelectVC further defines a value scoring function $\phi(\mathcal{S})$ that computes an outlier score for every single value in the full value set $\mathcal{V}$ by modeling the couplings between the single value and the values in the value subset $\mathcal{S}$. These two functions are iteratively reinforced until a stationary $\mathbf{q}$ is found. After obtaining value outlierness, given an object $\mathbf{x}$, we can integrate the outlierness of values contained by $\mathbf{x}$ to compute the object outlierness.

SelectVC is fundamentally different from existing frameworks in that: (i) SelectVC models the interactions with only the outlying behaviors. This avoids the interference from irrelevant couplings between irrelevant features, which significantly challenge full space-based approaches; and (ii) SelectVC unifies the two dependent tasks, value selection and outlier scoring, to optimize its outlier scoring, while existing subspace/feature selection-based approaches separate subspace/feature selection from outlier scoring and thus the subspaces/features retained by subspace/feature selection may be irrelevant to subsequent outlier detectors.

3.1 Value Subset Evaluation Function ψ

Since SelectVC aims to capture interactions of a value with only outlying values, function ψ is required to select a value subset $\mathcal{S}$ that consists of the most likely outlying values to facilitate the value outlier scoring in the next stage.

Definition 3.1 (Value Selection). *Value subset evaluation function ψ is to select a value subset $\mathcal{S}$ that contains the most likely outlying values from all the possible $\binom{|\mathcal{V}|}{|\mathcal{S}|}$ subsets.*

The value selection here is similar as feature selection, but we work on the value level. Nevertheless, subset search methods for feature selection, such as sequential search, random search and complete search [18], can be used to select a proper value subset.

3.2 Selective Value Coupling-based Scoring Function ϕ

Outlying behaviors are often strongly bond together while they are weakly coupled with other behaviors [20]. For example, the abnormal symptoms of diseases (e.g., the suspected signs like frequent urination, tiredness, and excessive thirsty for diabetes) are often concurrent, whereas they have weak association with normal symptoms or misdiagnosed abnormal symptoms.

SelectVC exploits such contrasting couplings to compute value outlierness by modeling the selective value couplings with only the outlying value set $\mathcal{S}$.

Definition 3.2 (Value Scoring). *The value scoring function $\phi : \mathcal{V} \mapsto \mathbb{R}$ exploits the couplings of a given value $v \in \mathcal{V}$ with the value subset $\mathcal{S}$ to compute the outlierness of the value v:*

$$\mathbf{q}(v) = \phi_v(\mathcal{S}) = \odot_{s \in \mathcal{S}} \eta(v, s), \tag{1}$$

where $\eta(\cdot, \cdot)$ captures the relation between the two values v and s, e.g., joint probability and conditional probability, and $\odot$ denotes one type of integration over η, e.g., first-order linear (or polynomial non-linear) summation and multiplication.

By working on the selective value couplings, SelectVC minimizes the interference from irrelevant features while captures the sufficient relevant information to assign larger outlierness to outlying values than normal/noisy values.

3.3 Stationary Criterion

The total number of possible value subsets is huge and different value subsets will result in very different value outlierness vectors. SelectVC aims to produce a stationary value outlierness vector to facilitate stable outier detection performance. Since we evaluate the convergence w.r.t. a vector, widely-used vector norms can be used. Let t be the iteration number, then a p-norm-based stationary criterion can be defined as follow.

$$\lim_{t \to \infty} ||\mathbf{q}_{t+1} - \mathbf{q}_t||_p \leq \epsilon, \tag{2}$$

where $p \geq 1$ and ϵ is a small constant.

4 THE SelectVC INSTANCE: POP

The SelectVC framework can be instantiated by specifying its three components: value scoring function ϕ, value subset evaluation function ψ, and the stationary criterion. The POP instance specifies these three components as follows. POP first specifies the functions ψ and ϕ by a top-k value selection function and a partial outlierness propagation-based value scoring function, respectively. POP then defines a stationary criterion using ℓ_1-norm.

4.1 Specifying ψ Using Top-k Outlying Value Selection

Given a value outlierness vector $\mathbf{q}$, POP defines a top-k outlying value selection function to select a value subset $\mathcal{S}$ containing a k *proportion* of the most outlying values from the full value set $\mathcal{V}$.

Definition 4.1 (Top-k Outlying Value Selection). *The top-k outlying value selection selects a value subset $\mathcal{S}$ with the cardinality $k|\mathcal{V}|$ from the full value set $\mathcal{V}$ as follows.*

$$\psi(\mathbf{q}) = \underset{\mathcal{S}\subset\mathcal{V} \text{ and } |\mathcal{S}|=k|\mathcal{V}|}{\arg\max} \sum_{s\in\mathcal{S}} \mathbf{q}(s). \tag{3}$$

Since $\mathbf{q}$ contains the outlierness of all feature values, after using the entries in $\mathbf{q}$ to sort the values in a descending order, Equation (3) is equivalent to selecting the top-ranked $k|\mathcal{V}|$ values. This value selection can be done in linear time, which well guarantees the scalability of POP to very high-dimensional data.

Note that outlying value selection is nontrivial due to the presence of noisy values and the huge search space. Simply selecting the most infrequent values may include the noisy values and consequently downgrade the quality of value outlierness estimation. Therefore, in the next section, POP initializes the value selection based on the frequencies of individual values but jointly optimizes the value selection and value scoring to obtain reliable outlying value sets and value outlierness.

4.2 Specifying ϕ by Partial Outlierness Propagation

POP defines a partial outlierness propagation-based function ϕ to leverage the contrasting couplings between outlying values to the selected subset $\mathcal{S}$ and normal/noisy values to the subset $\mathcal{S}$.

POP first builds a $|\mathcal{V}| \times |\mathcal{S}|$ matrix to capture the selective couplings of the values in the full value set $\mathcal{V}$ with the values in $\mathcal{S}$ using conditional probability.

Definition 4.2 (Selective Coupling Matrix). *The relation between the values in $\mathcal{V}$ and the values in $\mathcal{S}$ is captured by the selective coupling matrix $\mathbf{M} \in \mathbb{R}^{|\mathcal{V}|\times|\mathcal{S}|}$ which is defined as:*

$$\mathbf{M} = \begin{bmatrix} \eta(v_1, s_1) & \dots & \eta(v_1, s_{|\mathcal{S}|}) \\ \vdots & \ddots & \vdots \\ \eta(v_{|\mathcal{V}|}, s_1) & \dots & \eta(v_{|\mathcal{V}|}, s_{|\mathcal{S}|}) \end{bmatrix}, \; v_i \in \mathcal{V}, s_j \in \mathcal{S}, \tag{4}$$

where $\eta(v_i, s_j) = P(s_j|v_i) = \frac{freq(v_i, s_j)}{freq(v_i)} \in [0, 1]$ and freq denotes a frequency counting function.

Let u and u' be outlying and normal values, respectively. Given an outlying values $s \in \mathcal{S}$, since outlying values are often concurrent and the co-occurrence frequency is upper bounded by the frequency of s, we often have $freq(u, s) \gtrapprox freq(u', s)$. Moreover, per definition of outliers, $freq(u) \ll freq(u')$. Therefore, we normally obtain $\eta(u, s) > \eta(u', s)$ or $\eta(u, s) \gg \eta(u', s)$.

Let u'' be a noisy value. We may assume $freq(u) \approx freq(u'')$ as both u and u'' are infrequent. Since noisy values and outlying values are mainly contained by normal objects and outliers, respectively, u'' is presumed to have lower joint probabilities with the outlying values in $\mathcal{S}$, compared to the outlying value u. Thus, we also obtain $\eta(u, s) > \eta(u'', s)$. This demonstrates that the inherent asymmetrical property of conditional probability enables POP to effectively capture the aforementioned contrasting couplings.

POP further defines a partial outlierness propagation-based value scoring function ϕ by using $\mathbf{M}$ to propagate the outlierness of values in $\mathcal{S}$ to influence the scoring of values in $\mathcal{V}$.

Definition 4.3 (Partial Outlierness Propagation-based Value Scoring). *The partial outlierness propagation-based value scoring function ϕ is defined as follows.*

$$\mathbf{q}_{t+1}(v) = \phi_v(\mathcal{S}_t) = \sum_{s\in\mathcal{S}_t} \tilde{\mathbf{M}}(v, s)\mathbf{q}_t(s), \tag{5}$$

where $\tilde{\mathbf{M}}$ denotes a column-wise normalization of $\mathbf{M}$, $\mathbf{q}_t$ is normalized into a ℓ_1-norm unit, and $t \in \mathbb{Z}^+$ is a positive integer.

Equation (5) models the selective value couplings by simulating to partially propagating the t-th step value outlierness to the outlierness scoring in the $(t + 1)$-th step. Such partial outlierness propagation assits POP to iteratively enlarge the outlierness gap between the top-ranked values and the rest of values in the outlierness vector $\mathbf{q}$.

We initialize the vector $\mathbf{q}$ as follows.

$$\mathbf{q}_1(v) = \frac{freq(m) - freq(v)}{freq(m)} + \frac{freq(b) - freq(m)}{freq(b)}, \; v \in \mathcal{V}, \tag{6}$$

where m is the mode (i.e., the value occurs most frequently in a feature) of the feature containing the value v, and b is the benchmark value that has the largest frequency over all the values in $\mathcal{V}$.

This initialization is essentially built on the frequencies of individual values. Taking account of the location parameter (i.e., the mode) of the frequency distributions in Equation (6) is to produce a good initialization when the frequency distributions are very skewed across the features.

4.3 ℓ_1-Norm Stationary Criterion

A ℓ_1-norm-based stationary criterion is used in POP.

Definition 4.4 (ℓ_1-Norm Stationary Criterion). *A value outlierness vector $\mathbf{q}$ is stationary when satisfying:*

$$\Delta = ||\mathbf{q}_{t+1} - \mathbf{q}_t||_1 = \sum_{v\in\mathcal{V}} |\mathbf{q}_{t+1}(v) - \mathbf{q}_t(v)| \leq \epsilon, \tag{7}$$

where $\epsilon = 10^{-4}$ is used.

Actually, since the matrix $\mathbf{M}$ is fixed, POP obtains the stationary status when the values and their ranks in $\mathcal{S}$ do not change.

4.4 The Algorithm and Its Time Complexity

Algorithm 1 presents the procedures of detecting outliers using POP. Steps (1-6) are performed to obtain a $|\mathcal{V}| \times |\mathcal{V}|$ *full value* *coupling matrix* $\mathbf{M}'$. Since the conditional probabilities are fixed for all value pairs, we generate $\mathbf{M}'$ to facilitate quick access to the selective coupling matrix $\mathbf{M}$, which avoids re-scanning the data in the later iteration. Steps (7-11) performs the joint value selection and value scoring process to obtain the stationary $\mathbf{q}$. After obtaining the value outlierness, we compute the outlierness of data objects in Steps (13-15). In Step (14), following [20], we compute the outlierness of an object $\mathbf{x}$ as the weighted outlierness summation of its values, in which x_f denotes the value of $\mathbf{x}$ in feature f and $\omega_f = \sum_{v \in dom(f)} \mathbf{q}(v)$. Such weighted outlierness integration highlights relevant features and facilitates a proper object outlierness estimation. An object outlierness ranking R is finally returned in Step (17). The top-ranked objects in R are the most likely outliers.

Algorithm 1 *POP-based Outlier Detection*

```
Input: $\mathcal{X}$ - data objects, $k$ - a proportion of the full value set
Output: $R$ - an outlier ranking
 1: Initialize a $|\mathcal{V}| \times |\mathcal{V}|$ matrix $\mathbf{M}'$ for full value couplings
 2: for $v$ in $\mathcal{V}$ do
 3:     for $v'$ in $\mathcal{V}$ do
 4:         $\mathbf{M}'(v, v') \leftarrow \frac{freq(v, v')}{freq(v)}$
 5:     end for
 6: end for
 7: Initialize $\mathbf{q} \in \mathbb{R}^{|\mathcal{V}|}$ using Equation (6)
 8: repeat
 9:     $\mathcal{S} \leftarrow \underset{\mathcal{S} \subset \mathcal{V} \text{ and } |\mathcal{S}| = k|\mathcal{V}|}{\arg\max} \sum_{s \in \mathcal{S}} \mathbf{q}(s)$
10:     $\mathbf{q} \leftarrow \tilde{\mathbf{M}}_{|\mathcal{V}| \times |\mathcal{S}|} \times \mathbf{q}_{|\mathcal{S}| \times 1}(\mathcal{S})$
11: until Converge or reach the maximum iteration 200
12: Initialize $\mathbf{r} \in \mathbb{R}^{|\mathcal{X}|}$ as an outlierness vector for data objects
13: for $\mathbf{x}$ in $\mathcal{X}$ do
14:     $\mathbf{r}(\mathbf{x}) \leftarrow \sum_{f \in \mathcal{F}} \mathbf{q}^*(x_f)\omega_f$
15: end for
16: $R \leftarrow$ Sort $\mathcal{X}$ w.r.t. $\mathbf{r}$ in descending order
17: return $R$
```

POP requires one scanning over the data objects to obtain $\mathbf{M}'$ in Steps (1-6), which has $O(|\mathcal{X}||\mathcal{V}|^2)$. The iterations in Steps (8-11) have $O(|\mathcal{V}||\mathcal{S}|)$ time complexity. The object outlierness scoring and sorting take $O(|\mathcal{X}||\mathcal{V}|)$ in Steps (12-16). Therefore, the overall time complexity of POP is linear w.r.t. the data size and quadratic w.r.t. the total number of values. Since the average number of values per feature is normally very small, POP also has quadratic time complexity w.r.t. the number of features.

5 THEORETICAL ANALYSIS OF POP

This section analyzes the quality of the vector $\mathbf{q}^*$, the capability of POP in handling high-dimensional data and the setting of k.

5.1 Quality of the Stationary Vector $\mathbf{q}^*$

We show below that $\mathbf{q}$ becomes stable when the values in the selected subset $\mathcal{S}$ have the largest total pointwise mutual information.

Theorem 5.1 (Stationary Vector). *Let pmi($\mathcal{U}$) be the total pointwise mutual information among the values in a value set $\mathcal{U}$, i.e., pmi($\mathcal{U}$) = $\sum_{u \in \mathcal{U}} \sum_{u' \in \mathcal{U}} \log \frac{P(u,u')}{P(u)P(u')}$. Then, the value outlierness vector* $\mathbf{q}$ *converges to a vector* $\mathbf{q}^*$ *s.t.* $\forall \mathcal{U} \subseteq \mathcal{V}$ *and* $|\mathcal{U}| = |\mathcal{S}^*|$, *pmi($\mathcal{S}^*$) ≥ pmi($\mathcal{U}$), where $\mathcal{S}^*$ is the stationary value subset.*

Proof. At each iteration of POP, the subset $\mathcal{S}$ is updated until convergence, while the value conditional probability matrix $\mathbf{M}$ is fixed. Therefore, $\mathbf{q}$ becomes stationary when $\mathcal{S}$ does not change, i.e., $||\mathbf{q}_{t+1} - \mathbf{q}_t||_1 \leq \epsilon$ if $\mathcal{S}_t \subseteq \mathcal{S}_{t+1}$ and $\mathcal{S}_{t+1} \subseteq \mathcal{S}_t$.

Since $\mathbf{q}$ is updated using the conditional probabilities of a given value $v \in \mathcal{V}$ on the value subset $\mathcal{S}$, $\mathbf{q}(v)$ is primarily determined by the probabilities of the values in $\mathcal{S}$ given value v. Therefore, $\mathbf{q}(v) \propto \sum_{s \in \mathcal{S}} P(s|v)$ and thus $\mathbf{q}(\mathcal{S}) = \sum_{s' \in \mathcal{S}} \mathbf{q}(s') \propto \sum_{s' \in \mathcal{S}} \sum_{s \in \mathcal{S}} P(s|s')$. We have $\mathbf{q}(\mathcal{S}) \propto \sum_{s' \in \mathcal{S}} \sum_{s \in \mathcal{S}} \frac{P(s|s')}{P(s)}$ after taking account of the way we initialize $\mathbf{q}$. We will obtain a value subset $\mathcal{S}^*$ which has the largest *pmi* by maximizing $\mathbf{q}(\mathcal{S})$, and subsequently obtain $\mathbf{q}^*$ based on the subset $\mathcal{S}^*$. $\mathcal{S}^*$ remains unchanged since $\psi(\mathcal{S}^*)$ is already maximized, and thus $\mathbf{q}^*$ becomes stationary.

□

It is well known in natural language processing that pointwise mutual information biases towards rare words [25], i.e., pointwise mutual information between concurrent rare words are generally much larger than commonly-used or frequent words. In our case, this implies that the top-ranked values in the stationary vector $\mathbf{q}^*$ are normally outlying values - values which are exceptionally rare and have mutual interactions. In other words, POP can often obtain a highly discriminate outlierness vector where outlying values have larger outlierness than normal and noisy values.

5.2 Handling Distance Concentration Effect

The *concentration of distances* is a major issue in the curse of dimensionality. The distance concentration effect states that the discrimination between the near and far neighbors of a data object diminishes with increasing dimensions, in particular when the increased dimensions are irrelevant features [26].

Since we focus on value outlierness estimation, in general, we expect $||\mathbf{q}_t(u) - \mathbf{q}_t(v)||_p$ to be sufficiently large if u and v are respective outlying values and normal/noisy values, and to be small otherwise. Let v to be a normal value, without loss of generality, there exists a normal value u as its nearest neighbor and an outlying value w as its farthest neighbor. For a given value set $\mathcal{U} \subseteq \mathcal{V}$, according to the concentration effect theory [26], however, we have

$$\lim_{|\mathcal{U}| \to \infty} \frac{max\_d - min\_d}{min\_d} = 0, \tag{8}$$

$max\_d = || \sum_{w' \in \mathcal{U}} \tilde{\mathbf{M}}'(v, w')\mathbf{q}'_t(w') - \sum_{w' \in \mathcal{U}} \tilde{\mathbf{M}}'(w, w')\mathbf{q}'_t(w')||_p$ and $min\_d = || \sum_{w' \in \mathcal{U}} \tilde{\mathbf{M}}'(v, w')\mathbf{q}'_t(w') - \sum_{w' \in \mathcal{U}} \tilde{\mathbf{M}}'(u, w')\mathbf{q}'_t(w')||_p$ denote the largest and smallest distances to v, respectively.

As shown in [26], the concentration effect becomes more and more severe as the number of irrelevant features increases. Therefore, the larger the size of the value subset $\mathcal{U}$ is, we would be likely to have more severe concentration effect. The concentration effect is maximal when we use the full value couplings, i.e., to set $\mathcal{U} = \mathcal{V}$. POP substantially reduces such effect by working on a small value subset. POP could well overcome the concentration effect when setting k to be a sufficiently small value, but POP may lose relevant value couplings when k is too small. We will provide a general guideline for setting k in the next section.

5.3 Guidelines for Setting k

This section provides some guidelines for tuning the only parameter k, in particular for high-dimensional and small-sized data, based on three observations that (i) outliers typically account for only a small proportion of a data set; (ii) outliers often demonstrate their exceptional behaviors in only a small feature subset in high-dimensional data; and (iii) large k may lead to more severe distance concentration effect.

Theorem 5.2 (Maximum Number of Outlying Values). *Let O be the set of outlier objects in the data set $\mathcal{X}$, I be the maximum number of outlying values contained by an outlier $\mathbf{o} \in O$, and H be the total number of all possible outlying values in $\mathcal{X}$. Then*

$$H \leq I|O|. \tag{9}$$

Proof. When all outliers in O manifest different outlying values, we have $H = I|O|$. If there exists at least one $\mathbf{o} \in O$ sharing the same outlying values with other outliers, then $H < I|O|$. □

Corollary 5.2.1 (Upper Bound for k). *Let $\mathcal{S}^*$ be the value subset containing exactly all the possible outlying values, i.e., $|\mathcal{S}^*| = H$ and $k^* = \frac{|\mathcal{S}^*|}{|\mathcal{V}|}$. In high-dimensional and small-size data, i.e., $|\mathcal{V}| > |\mathcal{F}| > |\mathcal{X}|$, we have*

$$k^* \leq \frac{I|O|}{|\mathcal{V}|} < \frac{I|O|}{|\mathcal{X}|}. \tag{10}$$

According to Corollary 5.2.1, k^* is upper bounded by the outlier proportion $\frac{|O|}{|\mathcal{X}|}$ and the number of outlying values contained per outlier I in a high-dimensional and small-size data set. In general, $k^* < 0.5$ is a good bound based on the above three observations. Since our goal is to select a reliable outlying value subset and to substantially reduce the concentration effect, $k < k^*$ is suggested. We show in Section 6.8 that POP with $k = 0.3$ obtains stable performance in data sets with diverse dimensions.

6 EXPERIMENTS AND EVALUATION

We perform experiments to answer the following six questions:

- **Q1. Effectiveness in real-world data.** How accurately does POP detect outliers in real-world high-dimensional data with different levels of irrelevant features?
- **Q2. Significance of partial outlierness propagation.** How well does partial outlierness propagation perform compared to full outlierness propagation?
- **Q3. Significance of joint value selection and outlier scoring.** Can we replace POP with two independent successive modules: feature selection and outlier detection?
- **Q4. Scalability.** Does POP have good scalability?
- **Q5. Sensitivity.** How sensitive is POP to k?
- **Q6. Convergence.** How fast does POP converge?

6.1 Experiment Environment

POP and its competitors are implemented in JAVA. The implementations of all the competitors are obtained from their authors or the open-source platform ELKI [1]. All the experiments are executed at a node in a 3.4GHz Titan Cluster with 96GB memory.

6.2 Performance Evaluation Methods

All the outlier detectors finally produce an object ranking based on the outlier scores of the objects, i.e., the top-ranked objects are the most likely outliers. We measure the quality of the ranking by the area under ROC curve (AUC) which is computed by *Mann-Whitney-Wilcoxon* test [12]. AUC is one of the most popular performance evaluation methods and it inherently takes account of the class-imbalance nature, making the AUC results comparable across different data sets [8]. AUC ranges from zero to one. Higher AUC indicates better accuracy. The AUC value would be close to 0.5 given a random ranking of data objects. The *Wilcoxon* signed rank test is used to examine the significance of the AUC performance of POP against its competitors.

Data indicator refers to measures that capture inherent characteristics of data sets. These measures are strongly correlated with the performance of outlier detectors. Two data indicators, *coupling strength* (*coup*) and *outlier separability* (*sep*), are defined to assess the complexity of the data sets. They are briefly introduced below, and their quantization is reported in Table 1.

- *coup* represents the coupling strength between the outlier class label and its associated values. We use the probability of the outlier label given a single feature value to measure their coupling strength. $coup_{\mathcal{U}}$ is defined as the average conditional probability of the outlier class label over all its values in a value set $\mathcal{U}$, i.e., $coup_{\mathcal{U}} = \frac{1}{|\mathcal{U}|}\sum_{u \in \mathcal{U}} P(outlier|u)$. High $coup_{\mathcal{U}}$ indicates strong couplings between the outlier class and the values in $\mathcal{U}$.
- *sep* describes the difficulty in separating outliers from normal objects. *Feature efficiency* is a widely used indicator for measuring class separability in classification [14], which is referred to as the capability of a feature in enabling classifiers to make correct classification. We define the efficiency of a feature for outlier detection by the AUC performance of using the value marginal probabilities to identify outliers on the feature. *sep* is the maximum feature efficiency. A data set having a high *sep* indicates that the data set contains at least one highly relevant features.

6.3 Data Sets

Twelve publicly available real-world data sets [1] are used, which cover diverse domains, e.g., Internet advertising, image object recognition, web page classification and text classification, as shown in Table 1. Following the literature (e.g., [3, 8, 19–21]), eight of these data sets are directly transformed from highly imbalanced classification data, where the smallest class is treated as outliers and the largest class is normal; and we transform the other four balanced data sets (*PCMAC*, *BASE*, *WebKB*, *RELA*) by randomly sampling a small subset of the smallest class as outliers and keeping the largest class as normal class, such that the newly created data sets contain 2% outliers. The performance of these downsampled data sets is taken average over 10 times sampling. These transformation methods guarantee that the outlier class chosen is either a rare

[1] The used data sets are available at http://archive.ics.uci.edu/ml, http://featureselection.asu.edu, https://people.cs.umass.edu/~marlin/data.shtml, http://mmlab.ie.cuhk.edu.hk/projects/CelebA.html, http://tunedit.org/repo/data/text-wc, or https://www.csie.ntu.edu.tw/~cjlin/libsvmtools/.

class or a class with outlying semantics. All data sets are used with categorical features only. Features containing only one value are removed as they contain no information for outlier detection.

6.4 Q1. Effectiveness in Real-world Data

6.4.1 Experimental Settings. POP is compared with five detectors: CBRW [20], ZERO [22], iForest [19], ABOD [16] and LOF [7] on the 12 real-world data sets to evaluate its effectiveness.

- *Subspace-based Competitors*: ZERO and iForest. Both ZERO and iForest are state-of-the-art non-deterministic subspace methods [2]. Their performance is taken average from 10 runs. iForest and ZERO are used with the recommended settings in [19, 22], respectively.
- *Full Space-based Competitors*: CBRW, ABOD and LOF. CBRW is a state-of-the-art outlier detector for categorical data and it is closely related to POP. ABOD is an angle-based method which is specially designed for high-dimensional data. LOF is one of the most popular methods that works on full dimensionality and it is used as a baseline competitor. As recommended in [20], $\alpha = 0.95$ is used in CBRW. ABOD is parameter-free. For LOF, small values are suggested for the neighborhood size *MinPts* in [7]. We performed LOF with a range of different *MinPts*, i.e., $\{1, 5, 10, 20, 40, 60, 80, 100\}$, and report the results with $MinPts = 5$ as LOF using $MinPts = 5$ performs more stably across the data sets.

POP uses $k = 0.3$ by default. We will compare POP with feature selection-enabled methods in Section 6.6. Note that categorical data is transformed into numeric data to allow iForest, ABOD and LOF to work on the same data. The data sets are transformed by using a commonly used method 1-of-l (or one-hot) encoding [8, 22].

6.4.2 Findings - POP Performing Significantly Better Than Five State-of-the-art Outlier Detectors on Real-world High-dimensional Data. The AUC performance of POP and its five competitors: CBRW, ZERO, iForest, ABOD and LOF is reported in Table 1. POP performs better than all its five competitors on nine data sets, and significantly outperforms them at the 95% confidence level. On average, POP obtains more than 10%, 18%, 26%, 25% and 39% improvement over CBRW, ZERO, iForest, ABOD and LOF, respectively.

The data indicators $coup_{\mathcal{V}}$ and $coup_{\mathcal{S}}$ describe the coupling strength of the outlier class with the values in $\mathcal{V}$ and the values in $\mathcal{S}$, respectively. $livc = \frac{coup_{\mathcal{S}} - coup_{\mathcal{V}}}{coup_{\mathcal{V}}}$ therefore captures the level of *irrelevant value couplings* composed by the intersection of irrelevant value sets and the full value set. *livc* is a fine-grained value-level indicator which also implies the amount of irrelevant features per data. Higher *livc* indicates a larger percentage of irrelevant features a data set may contain. *livc* is used below to further explore the performance of these six detectors in data sets with different levels of irrelevant value couplings (or irrelevant features).

(1) Handling Data Sets with High livc. POP obtains the best performance on all the eight data sets with high *livc* (e.g. *livc* > 90%) (i.e., *w7a*, *wap.wc*, *R8*, *CAL16*, *AD*, *CAL28* , *CelebA* and *PCMAC*), and it averagely achieves substantial AUC improvement over its five competitors CBRW, ZERO, iForest, ABOD, and LOF by more than 13%, 21%, 30%, 24%, and 66%, respectively.

The superiority of POP is mainly because POP computes the outlier scores based on only selective (relevant) value interactions, which substantially improves the resilience of POP to irrelevant value couplings. LOF performs poorly on all these data sets due to two major reasons: (i) the severe distance concentration effect caused by the presence of a large amount of irrelevant features and (ii) the heavy dependency on an optimal neighborhood size *MinPts*, which varies substantially in data with different data sizes and data distributions. Compared to LOF, the competitors ABOD, ZERO and iForest are less sensitive to the irrelevant couplings, as they use more robust measures to define outlierness (e.g., angle between data objects) or work on feature subspaces. CBRW models complex value couplings to enlarge the outlier score difference between outlying values and other values, which enables CBRW to obtain significant improvements over the other four competitors. Nevertheless, CBRW still works on the full value couplings, and its performance is significantly downgraded by the irrelevant couplings compared to POP.

It is interesting that the methods like CBRW, ZERO, iForest and ABOD can obtain very good AUC performance in some data sets with many irrelevant couplings, e.g., *CAL16* and *CAL28*. This may be due to their high outlier separability, e.g., *CAL16* with $sep = 0.9613$ and *CAL28* with $sep = 0.9780$. In other words, these data sets contain some highly relevant features which, to some extent, enable these methods to address the noise brought by irrelevant features.

(2) Handling Data Sets with Low livc. As for the rest of the four data sets with low *livc*, i.e., *BASE*, *WebKB*, *RELA* and *Arrhy*, POP obtains the best performance on one data set, with two close to the best (having the difference in AUC less than 0.02), which is comparable to the best performer LOF. This is understandable since POP may omit some relevant value couplings when data sets have only limited irrelevant couplings, whereas LOF works on the full value interactions and thus captures the relevant couplings better.

It is interesting that all outlier detectors obtain quite small AUC values on these four data sets. This may be because all the four data sets have rather low outlier separability, as shown by the indicator *sep* in Table 1, and it is very challenging for learning methods to perform well on data sets without highly relevant features.

6.5 Q2. Significance of Partial Outlierness Propagation

6.5.1 Experimental Settings. POP is compared with its extreme variant called POP$^+$ which simulates full outlierness propagation by setting $k = 1.0$ to evaluate the significance of partial outlierness propagation in POP. Specifically, POP$^+$ computes value outlierness by $\mathbf{q}_{t+1}(v) = \sum_{u \in \mathcal{V}} \tilde{\mathbf{M}}'(v, u)\mathbf{q}_t(u)$, where $\mathbf{M}'$ is a $|\mathcal{V}| \times |\mathcal{V}|$ full value coupling matrix and $\tilde{\mathbf{M}}'$ is its column-wise normalization. Therefore, POP$^+$ is exactly the same as POP except that it uses the full value set $\mathcal{V}$ rather than the value subset $\mathcal{S}$ in POP.

6.5.2 Findings - POP Using Partial Outlierness Propagation Significantly Outperforming Its Counterpart Using Full Outlierness Propagation. The AUC performance of POP and POP$^+$ is reported in

[2] The computational time of deterministic subspace methods like FPOF [13] and Comprex [3] is prohibitive for high-dimensional data, and they run out of memory or cannot output the results for most of the used data sets within four weeks. Also, the empirical results in [20] show that CBRW significantly outperforms these methods. Thus, we focus on the comparison with CBRW and the other four competitors.

Table 1: A Summary of Data Sets Used, Indicator Quantization Results and AUC Performance of POP, POP$^+$ and Their Competitors: Five Full Space- or Subspace-based Outlier Detectors. $livc = \frac{coup_{\mathcal{S}} - coup_{\mathcal{V}}}{coup_{\mathcal{V}}}$ **describes the level of irrelevant value couplings per data. The middle horizontal line roughly separates data sets with high** *livc* **from that with low** *livc*. **CBRW runs out of memory on high-dimensional data** ***R8*** **and** ***WebKB*****. ABOD runs out-of-memory on large data** ***w7a*** **and** ***CelebA*****.**

Data Summary				Data Indicators				Our Methods		Competitors				
Data	Acronym	$\lvert\mathcal{X}\rvert$	$\lvert\mathcal{F}\rvert$	$coup_{\mathcal{V}}$	$coup_{\mathcal{S}}$	*livc*	*sep*	POP	POP$^+$	CBRW	ZERO	iForest	ABOD	LOF
w7a	-	49749	300	0.1490	0.4440	197.99%	0.5927	**0.8673**	0.8054	0.6460	0.5375	0.4053	NA	0.4996
wap.wc	-	346	4229	0.0306	0.0866	183.01%	0.9713	**1.0000**	0.9666	0.7900	0.6552	0.5558	0.5243	0.5161
Reuters8	R8	3974	9467	0.0358	0.0980	173.74%	0.9358	**0.9479**	0.9324	NA	0.8827	0.8443	0.7856	0.8916
Caltech-16	CAL16	829	253	0.1099	0.2961	169.43%	0.9613	0.9928	**0.9930**	0.9925	0.9878	0.9742	0.9766	0.3881
InternetAd	AD	3279	1555	0.1923	0.4370	127.25%	0.6982	**0.9290**	0.8300	0.7348	0.7062	0.7084	0.7023	0.5507
Caltech-28	CAL28	829	727	0.0654	0.1465	124.01%	0.9780	0.9608	**0.9616**	0.9599	0.9538	0.9377	0.9268	0.4390
CelebA	-	202599	39	0.0307	0.0665	116.61%	0.7961	0.8968	**0.8981**	0.8462	0.7595	0.6797	NA	0.4726
PCMAC	-	1002	3039	0.0327	0.0638	95.11%	0.7721	**0.6935**	0.6617	0.6332	0.5266	0.4767	0.4903	0.6198
BASEHOCK	BASE	1019	4320	0.0347	0.0613	76.66%	0.6292	0.6521	0.6329	0.6177	0.5287	0.4731	0.4883	**0.6639**
WebKB	-	1658	6601	0.0303	0.0526	73.60%	0.7501	0.7306	0.7266	NA	0.6950	0.6773	0.6701	**0.8250**
RELATHE	RELA	794	4080	0.0320	0.0554	73.13%	0.6365	**0.7449**	0.7173	0.7014	0.6047	0.5578	0.5685	0.7432
Arrhythmia	Arrhy	452	64	0.2548	0.4287	68.25%	0.6293	0.6762	0.6890	**0.6910**	0.6644	0.6868	0.5948	0.6008
							Average (Top-8)	**0.9110**	0.8811	0.8004	0.7512	0.6978	0.7343	0.5472
							Average (All)	**0.8410**	0.8179	0.7613	0.7085	0.6648	0.6728	0.6009
							P-value	-	0.0269	0.0098	0.0005	0.0010	0.0020	0.0122

Table 1. Although POP uses more than two-thirds less information than POP$^+$, it obtains about 3% improvement over POP$^+$ and significantly outperforms POP$^+$ at the 95% confidence level. POP outperforms POP$^+$ on eight data sets, with the maximal improvement up to 11%, and it performs very comparably to POP$^+$ on the other four data sets.

POP$^+$ works on the original data space which contains much more irrelevant value couplings than the clean data space that POP works on, as indicated by the substantial difference between $coup_{\mathcal{V}}$ and $coup_{\mathcal{S}}$ in Table 1. As a result, even though POP$^+$ is operated on the data space that contains the condensed data space used by POP, its performance is significantly degraded due to two major reasons: (i) its distance concentration effect is more severe and (ii) its full outlierness propagation amplifies irrelevant couplings and makes negative propagation.

Note that although POP$^+$ underperforms POP, it substantially outperforms all the five competitors in Table 1. This may explain that the (either partial or full) outlierness propagation mechanism well captures contrasting couplings between outlying-to-outlying values and normal/noisy-to-outlying values and has better capability in handling high-dimensional data than the five competitors.

6.6 Q3. Significance of Joint Value Selection and Outlier Scoring

6.6.1 Experimental Settings. There are two major ways to replace POP with two independent successive modules: feature selection and outlier detection, which are described as follows.

- The value subset selected by POP can be used to perform feature selection. That is, for each data set, we create a corresponding new data set with a subset of features spanned by the values in the selected value subset. We denote this feature selection method as POFS. The existing outlier detectors can then be performed on the newly created data.
- Alternatively, existing outlier detectors can be combined with previously proposed feature selection methods which are designed for outlier detection. Two of the latest outlying feature selection methods: CBRW_FS (denoted by CBFS) [20] and DSFS [21] are used. CBFS only returns a feature ranking. CBFS is aligned with POFS and selects the top-ranked $|\mathcal{F}'|$ features, where $\mathcal{F}'$ denotes the feature subset selected by POFS. DSFS outputs a feature subset $\mathcal{F}''$ without any parameters.

The five outlier detectors with the same settings described in Section 6.4 are used with POFS, CBFS and DSFS to have a comprehensive comparison to POP. This enables us to examine how critical it is for the joint process of value selection and outlier scoring, compared to perform feature/value selection and outlier detection independently.

6.6.2 Findings - Joint Value Selection and Outlier Scoring Enabling POP to Obtain More Than 5% Improvement Over the Best Performer Among All the Successive Combinations of Three Outlying Feature/Value Selection Methods and Five State-of-the-art Outlier Detectors. The AUC performance of POP and all the 15 combinations of the three feature selection methods POFS, CBFS and DSFS and the five detectors CBRW, ZERO, iForest, ABOD and LOF is reported in Table 2. The results show that POP significantly outperforms all the 15 combinations and obtains over 5% to 50% improvements.

The POFS or CBFS-empowered CBRW, ZERO, iForest and ABOD substantially improve the AUC performance over its original editions, but they still perform significantly less effectively than POP. This is due to two major reasons: (i) POFS or CBFS selects features independently from the these outlier detectors and thus the selected features are not optimal to these detectors, in contrast to POP in which value selection and value outlierness scoring function are simultaneously optimized; and (ii) POP works on value subsets whereas its competitors operates on feature subsets, so POP captures more fine-grained value interactions than its counterparts. All

Table 2: AUC Results of POP and the Combinations of the Five Competitors with Three Feature Selection Methods POFS, CBFS and DSFS on the 12 Data Sets. $|\mathcal{F}|$ denotes the number of original features, $|\mathcal{F}'|$ denotes the number of features retained by POFS and CBFS, and $|\mathcal{F}''|$ is the number of features retained by DSFS.

				POP	CBRW			ZERO			iForest			ABOD			LOF		
Data	$\|\mathcal{F}\|$	$\|\mathcal{F}'\|$	$\|\mathcal{F}''\|$	-	POFS	CBFS	DSFS	POFS	CBFS	DSFS	POFS	CBFS	DSFS	POFS	CBFS	DSFS	POFS	CBFS	DSFS
w7a	300	180	26	**0.8673**	0.8220	0.7738	0.5155	0.7701	0.7885	0.5155	0.5893	0.7674	0.5155	NA	NA	NA	0.5661	0.6108	0.5010
wap.wc	4229	2537	3570	**1.0000**	0.9026	0.8739	0.6387	0.7339	0.7429	0.5395	0.5902	0.6816	0.5121	0.5566	0.7355	0.5437	0.6065	0.7161	0.4856
R8	9467	5680	2006	**0.9479**	NA	NA	0.9249	0.8902	NA	0.8758	0.8370	NA	0.8426	0.8020	NA	0.7902	0.8772	NA	0.7252
CAL16	253	151	194	0.9928	0.9930	0.9928	**0.9931**	0.9910	0.9900	0.9903	0.9828	0.9824	0.9811	0.9922	0.9908	0.9920	0.4327	0.4428	0.2923
AD	1555	933	49	**0.9290**	0.7845	0.7456	0.7432	0.7547	0.7587	0.7428	0.7345	0.7723	0.7435	0.7298	0.7548	0.7495	0.5760	0.6652	0.5233
CAL28	727	436	564	**0.9608**	0.9603	0.9604	0.9599	0.9566	0.9584	0.9540	0.9488	0.9524	0.9421	0.9507	0.9526	0.9402	0.2247	0.2393	0.3345
CelebA	39	23	34	**0.8968**	0.8901	0.8818	0.8502	0.8519	0.8511	0.7722	0.8038	0.8213	0.6973	NA	NA	NA	0.5644	0.6051	0.5220
PCMAC	3039	1823	1256	**0.6935**	0.6759	0.6678	0.6413	0.5952	0.5793	0.4959	0.5509	0.5425	0.4745	0.5582	0.5511	0.4580	0.6605	0.6574	0.5988
BASE	4320	2592	1895	0.6521	0.6294	0.6558	0.5760	0.5396	0.5897	0.4375	0.5096	0.5417	0.4233	0.5117	0.5666	0.4086	0.6666	**0.6984**	0.6187
WebKB	6601	3960	3487	0.7306	0.7449	NA	0.7251	0.7377	NA	0.6995	0.7292	NA	0.6891	0.7369	NA	0.6712	0.4543	NA	**0.8246**
RELA	4080	2448	2101	**0.7449**	0.7256	0.7352	0.6984	0.6580	0.6793	0.5987	0.6268	0.6459	0.5844	0.6338	0.6582	0.5718	0.7141	0.7334	0.6965
Arrhy	64	38	13	**0.6762**	0.6095	0.6527	0.5625	0.6074	0.6540	0.5626	0.6065	0.6543	0.5624	0.5341	0.5814	0.5540	0.6004	0.6230	0.5534
			Average	**0.8410**	0.7943	0.7940	0.7357	0.7572	0.7592	0.6820	0.7091	0.7362	0.6640	0.7006	0.7240	0.6679	0.5786	0.5992	0.5563
			P-value	-	0.0098	0.0117	0.0010	0.0024	0.0020	0.0005	0.0005	0.0020	0.0005	0.0059	0.0078	0.0020	0.0010	0.0098	0.0024

three feature selection methods do not improve the performance of LOF. This is mainly because LOF needs to re-tune its neighborhood size *MinPts* to obtain desirable performance on the data sets with reduced feature sets due to its sensitivity to the data distribution.

6.7 Q4. Scalability

6.7.1 Experiment Settings. We examine the scalability of POP w.r.t. both of data size and dimensionality.

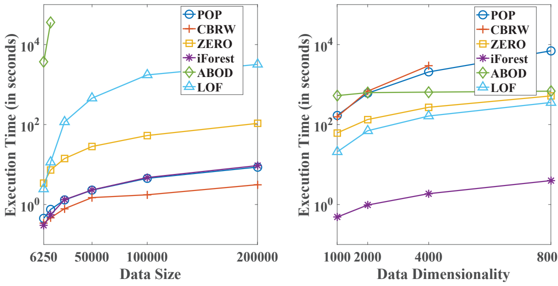

Figure 2: Scalability Test Results. ABOD and CBRW run out of memory when the number of objects reaches 25,000 and the number of features reaches 8,000, respectively.

We use six subsets of the largest data set *CelebA* to test the scalability w.r.t. data size. The smallest data subset contains 6,250 objects, and the sizes of subsequent subsets are increased by a factor of two until the largest subset containing 200,000 objects. All these data subsets contain the same number of features (i.e., 39).

In terms of scalability w.r.t. the number of features, four subsets of the data set with the largest number of features, *R8*, are used. The data subset with the lowest dimensionality contains 1,000 features, and subsequent data sets are created by increasing the dimensionality by a factor of 2, until the data set with highest dimensionality containing 8,000 features. All these four data subsets contain the same number of objects (i.e., 3,974).

6.7.2 Findings - POP Obtaining Good Scalability. As expected, POP is linear to the data size and quadratic to the number of features. In the left panel, POP runs comparably fast to CBRW, iForest and ZERO, and is two to four orders of magnitude faster than LOF and ABOD. In the right panel, POP and CBRW have similar runtime and they run considerably slower than the other four detectors, since both POP and CBRW model complex value interactions while the other four detectors ignore these interactions. Although POP and CBRW runs slower, they obtain significantly better AUC performance than their counterparts, as shown in Tables 1 - 2.

6.8 Q5. Sensitivity

6.8.1 Experimental Settings. We investigate the sensitivity of POP w.r.t. its only parameter k on all the 12 data sets using a wide range of k, i.e., $\{0.1, 0.2, 0.3, 0.4, 0.5\}$.

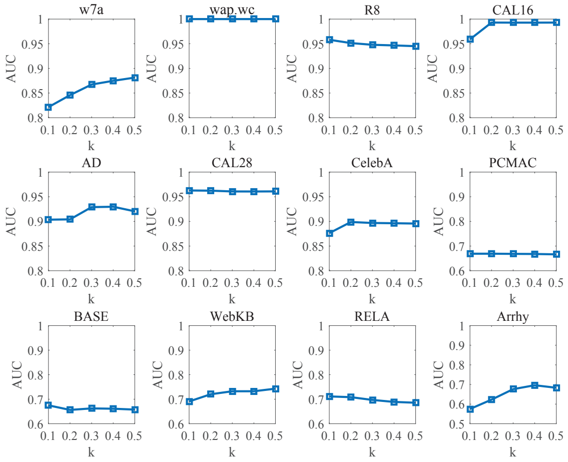

Figure 3: Sensitivity Test Results of POP w.r.t. *k*

6.8.2 Findings - POP Performing Stably w.r.t. k. The sensitivity test results of POP are shown in Figure 3. POP performs very stably w.r.t. k on all the data sets except *w7a* and *Arrhy* when k is chosen in $\{0.2, 0.3, 0.4\}$. This may be because POP is able to retain stable outlierness of the top-ranked outlying values in the value outlierness vector when the selected value subset mainly contains outlying values. We conjecture that the two data sets *w7a* and *Arrhy* may contain a larger proportion of outlying values, so a larger k is required to have a more effective modeling of the selective value couplings. In general, $k = 0.3$ is recommended in practice.

6.9 Q6. Convergence

6.9.1 Experimental Settings. This section examines the ℓ_1-norm convergence, i.e., $\Delta = ||\mathbf{q}_{t+1} - \mathbf{q}_t||_1$, on all the 12 data sets.

6.9.2 Findings - POP Obtaining Rapid Convergence. The convergence test results are presented in Figure 4. As expected, POP converges on all the 12 data sets. POP converges within 100 iterations in most of the data sets. POP takes slight longer time to converge in a few data sets, e.g., *w7a*, *BASE*, *WebKB* and *Arrhy*. This may be because these data sets contain larger percentages of outlying values, or they contain many noisy values that behave quite similarly as outlying values. Nevertheless, POP converges within 160 iterations on these data sets.

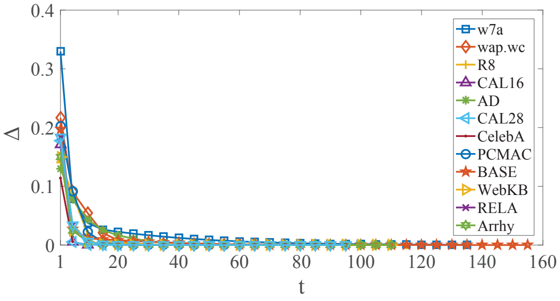

Figure 4: Convergence Test Results

7 CONCLUSIONS

A novel framework SelectVC is proposed to combine value selection with outlier scoring by iteratively learning selective value couplings to detect outliers in high-dimensional categorical data. SelectVC is further instantiated to a partial outlierness propagation-based method called POP. Our extensive empirical results show that (i) POP performs significantly better than 20 competitors, including five state-of-the-art full space- or subspace-based outlier detectors and their combinations with three outlying feature selection methods, on 12 real-world high-dimensional data with a variety of irrelevant features; (ii) The partial outlierness propagation enables POP to obtain about 3% AUC improvement, while the joint optimization enables POP to gain at least 5% AUC improvement; and (iii) POP obtains good scalability, stable performance w.r.t. the only parameter k and fast convergence rate. These results justify our key insight that modeling only selective value couplings enables us to well contrast outlying behaviors to non-outlying behaviors. In future, we plan to extend POP by capturing the interactions of values with a set of arbitrary-length outlying/normal patterns to identify more sophisticated outliers.

REFERENCES

[1] Elke Achtert, Hans-Peter Kriegel, Erich Schubert, and Arthur Zimek. 2013. Interactive data mining with 3D-parallel-coordinate-trees. In *SIGMOD*. 1009–1012.

[2] Charu Aggarwal and S Yu. 2005. An effective and efficient algorithm for high-dimensional outlier detection. *The VLDB Journal* 14, 2 (2005), 211–221.

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

[4] Fabrizio Angiulli, Fabio Fassetti, and Luigi Palopoli. 2009. Detecting outlying properties of exceptional objects. *ACM Transactions on Database Systems* 34, 1 (2009), 7.

[5] Fabrizio Angiulli and Clara Pizzuti. 2005. Outlier mining in large high-dimensional data sets. *IEEE Transactions on Knowledge and Data Engineering* 17, 2 (2005), 203–215.

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

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

[8] Guilherme O Campos, Arthur Zimek, Jörg Sander, Ricardo JGB Campello, Barbora Micenková, Erich Schubert, Ira Assent, and Michael E Houle. 2016. On the evaluation of unsupervised outlier detection: measures, datasets, and an empirical study. *Data Mining and Knowledge Discovery* 30, 4 (2016), 891–927.

[9] Longbing Cao. 2015. Coupling learning of complex interactions. *Information Processing & Management* 51, 2 (2015), 167–186.

[10] Ting Chen, Lu-An Tang, Yizhou Sun, Zhengzhang Chen, and Kai Zhang. 2016. Entity embedding-based anomaly detection for heterogeneous categorical events. In *IJCAI*. AAAI Press, 1396–1403.

[11] Amol Ghoting, Srinivasan Parthasarathy, and Matthew Eric Otey. 2006. Fast mining of distance-based outliers in high-dimensional datasets. In *SDM*. SIAM.

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

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

[14] Tin Kam Ho and Mitra Basu. 2002. Complexity measures of supervised classification problems. *IEEE Transactions on Pattern Analysis and Machine Intelligence* 24, 3 (2002), 289–300.

[15] Fabian Keller, Emmanuel Müller, and Klemens Bohm. 2012. HiCS: High contrast subspaces for density-based outlier ranking. In *ICDE*. IEEE, 1037–1048.

[16] Hans-Peter Kriegel and Arthur Zimek. 2008. Angle-based outlier detection in high-dimensional data. In *SIGKDD*. ACM, 444–452.

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

[18] Jundong Li, Kewei Cheng, Suhang Wang, Fred Morstatter, Robert P Trevino, Jiliang Tang, and Huan Liu. 2016. Feature Selection: A Data Perspective. *CoRR* abs/1601.07996 (2016).

[19] Fei Tony Liu, Kai Ming Ting, and Zhi-Hua Zhou. 2012. Isolation-based anomaly detection. *ACM Transactions on Knowledge Discovery from Data* 6, 1, Article 3 (2012), 39 pages.

[20] Guansong Pang, Longbing Cao, and Ling Chen. 2016. Outlier detection in complex categorical data by modelling the feature value couplings. In *IJCAI*. AAAI Press, 1902–1908.

[21] Guansong Pang, Longbing Cao, Ling Chen, and Huan Liu. 2016. Unsupervised feature selection for outlier detection by modelling hierarchical value-feature couplings. In *ICDM*. IEEE, 410–419.

[22] Guansong Pang, Kai Ming Ting, David Albrecht, and Huidong Jin. 2016. ZERO++: Harnessing the power of zero appearances to detect anomalies in large-scale data sets. *Journal of Artificial Intelligence Research* 57 (2016), 593–620.

[23] Ninh Pham and Rasmus Pagh. 2012. A near-linear time approximation algorithm for angle-based outlier detection in high-dimensional data. In *SIGKDD*. ACM, 877–885.

[24] Saket Sathe and Charu C Aggarwal. 2016. Subspace outlier detection in linear time with randomized hashing. In *ICDM*. IEEE, 459–468.

[25] Peter D Turney and Patrick Pantel. 2010. From frequency to meaning: Vector space models of semantics. *Journal of Artificial Intelligence Research* 37 (2010), 141–188.

[26] Arthur Zimek, Erich Schubert, and Hans-Peter Kriegel. 2012. A survey on unsupervised outlier detection in high-dimensional numerical data. *Statistical Analysis and Data Mining* 5, 5 (2012), 363–387.