Embedding-Based Complex Feature Value Coupling Learning for Detecting Outliers in Non-IID Categorical Data

Hongzuo Xu, Yongjun Wang, Zhiyue Wu,† Yijie Wang†*
†Science and Technology on Parallel and Distributed Processing Laboratory
College of Computer, National University of Defense Technology, China
{xuhongzuo13,wangyongjun,wangyijie}@nudt.edu.cn, zhiyue.wu@outlook.com

Abstract

Non-IID categorical data is ubiquitous and common in real-world applications. Learning various kinds of couplings has been proved to be a reliable measure when detecting outliers in such non-IID data. However, it is a critical yet challenging problem to model, represent, and utilise high-order complex value couplings. Existing outlier detection methods normally only focus on pairwise primary value couplings and fail to uncover real relations that hide in complex couplings, resulting in suboptimal and unstable performance. This paper introduces a novel unsupervised embedding-based complex value coupling learning framework EMAC and its instance SCAN to address these issues. SCAN first models primary value couplings. Then, coupling bias is defined to capture complex value couplings with different granularities and highlight the essence of outliers. An embedding method is performed on the value network constructed via biased value couplings, which further learns high-order complex value couplings and embeds these couplings into a value representation matrix. Bidirectional selective value coupling learning is proposed to show how to estimate value and object outlierness through value couplings. Substantial experiments show that SCAN (i) significantly outperforms five state-of-the-art outlier detection methods on thirteen real-world datasets; and (ii) has much better resilience to noise than its competitors.

Introduction

Outlier detection is the process of identifying rare and exceptional data objects that are dramatically different from the majority of data objects, which is important in many applications including intrusion detection, medical diagnose, and fraud detection. Non-independent and identically distributed (Non-IID) categorical data is ubiquitous and common in these real-world applications. However, it is still a challenging problem to detect outliers in such non-IID data with complex interactions and unavoidable noise.

Non-IID data poses following two major challenges: (i) Diversified frequency distributions across different features mean frequencies may have varying semantics. (ii) The sophisticated *couplings* (Cao, Ou, and Yu 2012) (i.e., different types and hierarchies of interactions) cannot be abstracted or weakened to the extent of satisfying the IIDness assumption of the most of the existing algorithms (Cao 2014).

In non-IID data, data objects and object features are not independent and identically distributed but coupled and personalised (Cao 2014). For example, in the task of detecting cancer patients given various of physical signs as features, cancer is related to multiple abnormal symptoms, such as persistent lumps, unexpected weight loss, and night sweats, which means the object features in real-world data are normally dependent or even exist relationships that are beyond dependency relation. On the other hand, different data objects have their own characteristics and personalities in real-world data, and the assumption that all the objects are identically distributed is often violated.

Besides the non-IIDness, another key complexity is the unavoidable noise in real-world data. Specifically, a dataset is often a mixture of relevant features and noisy features (i.e., features in which outliers behave normally while some normal objects are abnormal). These noisy features greatly blur the distinction between outliers and inliers. However, identification of these noisy features is non-trivial in categorical data since a feature may contain not only relevant values (i.e., values that can well indicate outliers) but also infrequent noisy values (i.e., values that randomly occur in normal objects).

Most of the early attempts for outlier detection in categorical data is based on the IIDness assumption, e.g., pattern-based methods like (He et al. 2005; Aggarwal and Yu 2005; Das, Schneider, and Neill 2008). Note that features in non-IID data have diversified frequency distributions. The generated patterns in these pattern-based methods are derived from different feature combinations, and thus the semantic and importance of pattern frequency differ significantly for different patterns (Pang, Cao, and Chen 2016). In addition, these methods are easily misled by noisy/irrelevant features because they are normally based on full space.

To detect outliers in such non-IID categorical data, coupling learning-based outlier detectors (Pang, Cao, and Chen 2016; Pang et al. 2017a; 2017b; Xu et al. 2018a; 2018b) are the major solution to capture the non-IIDness nature of real-world data. These methods model value couplings through various techniques to estimate outlierness. They can also partially resist the negative impact from noise, because investigating value interactions has shown to be an effective

*Corresponding author.

and reliable measure to distinguish relevant values and noisy values (Xu et al. 2018a). Nevertheless, these methods only successfully model pairwise primary value couplings and fail to uncover real relations that hide in high-order complex couplings. Therefore, they may perform ineffectively in complex non-IID data. It is still an extraordinarily challenging problem to model, represent, and utilise high-order complex couplings.

This paper also focuses on the coupling learning to handle non-IID categorical data. Non-IID data is often embedded with complex relationships (Zhou, Sun, and Li 2009; Cao 2014). Especially in the scenario of outlier detection, outliers often demonstrates outlierness through multiple behaviours (feature values), and thus these outlying behaviours are not independent but tend to be concurrent (Pang et al. 2017a; Xu et al. 2018a). Mining these homophily outlying-to-outlying value couplings is significant to discover outliers. However, real relations sometimes can only be discovered via high-order complex couplings. Note that we choose to investigate the most fine-grained value couplings rather than the couplings between data objects or features. It is because high-level couplings (couplings between objects and features) can be regarded as the integration of value couplings. In this paper, we wield an embedding method to learn and embed high-order complex value couplings into a value representation matrix. The challenging problem is how to drive the embedding method on categorical data, and meanwhile consider the essence of outliers. We construct a biased value coupling-based value network and employ a network embedding method to tackle this problem. Different kinds of relationships of the values in the network, which are originally represented by edges, structure characteristics, or other high-order topological measures of network, can be captured and encoded in the embedding vectors.

Based on the above basic concepts, a novel unsupervised Embedding-based coMplex vAlue Coupling learning framework (EMAC for short) is proposed for detecting outliers in non-IID categorical data. We use the EMAC framework to illustrate our insight of using embedding method to learn high-order complex value couplings. The specific implementation for each component of the framework can be replaced by different techniques.

EMAC framework is further instantiated to a method that learns complex value couplings by employing an extended Skip-gram architecture (i.e., *node2vec* (Grover and Leskovec 2016)) on a biased value Coupling-based vAlue Network (SCAN for short). SCAN is a specific implementation of EMAC framework. Specifically, SCAN models primary direct value couplings via Ochiai coefficient and conditional probability, and obtains indirect value couplings by employing cosine similarity. Value coupling bias is defined to capture value couplings with different granularities and take the essence of outliers into consideration. *node2vec* is performed on the biased value coupling-based value network to obtain the value representation matrix embedded with high-order complex value couplings. After getting reliable value couplings, we further propose value subspace-based Bidirectional Selective Value Coupling (BSVC) learning to evaluate value and object outlierness, which can effectively alleviate the interference from noise.

Accordingly, the main contributions of this paper are:

- We introduce a novel framework EMAC that employs value representation matrix to embed high-order complex value couplings, which provides a reliable insight for outlier detection in non-IID categorical data.
- EMAC is further instantiated to SCAN. SCAN provides a specific implementation to practically tackle the problem of how to model and represent complex value couplings. In addition, SCAN further introduces how to utilise complex value couplings to evaluate outlierness by proposing BSVC learning.

Substantial experiments show that SCAN (i) significantly outperforms five state-of-the-art methods on thirteen real-world datasets; (ii) has better resilience to noise than its competitors; and (iii) performs stably w.r.t. its parameters.

Related Work

Traditional outlier detection algorithms are normally based on IIDness assumption and identify outliers in original data space, e.g., (He et al. 2005; He, Deng, and Xu 2005; Aggarwal and Yu 2005; Das, Schneider, and Neill 2008; Akoglu et al. 2012; Wu and Wang 2013). They may fail to obtain effective performance in widespread real-world non-IID data. In addition, the results of these methods are often considerably biased by noisy/irrelevant features which distort the data by masking outliers as normal objects.

In order to handle noisy data, feature subspace-based methods, e.g., (Lazarevic and Kumar 2005; Keller, Müller, and Bohm 2012; Liu, Ting, and Zhou 2012; Sathe and Aggarwal 2016), are popularly proposed because outliers are usually embedded in locally relevant subspaces (Aggarwal 2017). However, a feature may contain not only relevant values but also noisy values. Comparing with value subspace-based methods, these coarse-grained methods fail to differentiate between relevant values and noisy values of the same feature, i.e., they may omit informative relevant values or mix noisy values when removing or retaining an entire feature (Xu et al. 2018a).

Non-IID learning has aroused increasing attention in recent years, e.g., (Cao 2014; Cao and Yu 2016; Chen et al. 2016; Cinbis, Verbeek, and Schmid 2016; Jian et al. 2018; Zhu et al. 2018; Zhao et al. 2018). Coupling-based methods are the major solution when detecting outliers in the non-IID categorical data. The method CBRW (Pang, Cao, and Chen 2016) models intra- and inter-feature value couplings to evaluate value outlierness. However, its performance can be greatly downgraded by the overwhelming noise because it is based on data full space. Feature selection-based method HOUR (Pang et al. 2017a) that employs a wrapper approach to iteratively optimise feature selection and outlier scoring is proposed to resist the negative impact from noisy features, whereas it is also hard for HOUR to differentiate between relevant values and noisy values of the same feature. To solve this problem, value subspace-based methods POP (Pang et al. 2017b), WOD (Xu et al. 2018b), and RHAC (Xu et al. 2018a) are proposed. POP (Pang et al. 2017b)

iteratively performs value selection and value scoring attempting to jointly optimise these two phases. WOD (Xu et al. 2018b) combines value clustering and weighted value coupling learning to evaluate value outlierness. RHAC (Xu et al. 2018a) introduces hierarchical value couplings. They can avoid the aforementioned disaster for feature subspace-based methods. Nevertheless, these existing coupling-based methods only model pairwise value couplings and may fail to capture real relations that hide in high-order complex value couplings. Thus, they may obtain suboptimal and unstable performance when they handle complex non-IID data.

EMAC for Learning Complex Value Couplings to Detect Outliers

The EMAC framework aims to generate a representation matrix embedded with complex and reliable value couplings for detecting outliers in non-IID categorical data. Let $\mathcal{X} = \{\boldsymbol{x}_1, \boldsymbol{x}_2, \cdots, \boldsymbol{x}_N\}$ be N data objects described by D categorical features $\mathcal{F} = \{f_1, f_2, \cdots, f_D\}$. $v_f^{\boldsymbol{x}}$ is the feature value of object $\boldsymbol{x}$ in feature f. The full set of feature values is the union of the distinct value domains from all the features, i.e.,$\mathcal{V} = \cup_{f \in \mathcal{F}} \mathcal{V}_f$, where $\mathcal{V}_f$ is the value domain of feature f and $\mathcal{V}_f \cap \mathcal{V}_{f'} = \emptyset, \forall f \neq f'$. Let $h : \mathcal{V} \rightarrow \mathbb{R}^r$ be the mapping function from values to numerical feature representations that we aim to learn for subsequent object outlierness evaluating. Here r is a parameter specifying the number of dimensions of value representation matrix.

In non-IID categorical data that exhibits abundant interactions between data objects, object features, and feature values, exploring couplings is necessary and unavoidable in different learning problems (Cao 2014). Especially, in outlier detection task, outliers often behaved abnormally through multiple outlying values, i.e., outlying values are concurrent and have strong interactions rather than independent (Pang et al. 2017a; Xu et al. 2018a). Thus, modeling reliable value relationships can facilitate the detecting of abnormal values. However, real value relationships sometimes can only be revealed through high-order complex value couplings.

It is a very challenging task to learn, represent, and utilise complex value couplings. Primary value couplings like conditional probability, mutual information, and various kinds of similarity measures can only evaluate pairwise value interactions but fail to capture high-order value couplings (e.g., deep transitive relation). Harnessing the power of embedding method, we propose EMAC framework to learn and represent high-order complex value couplings through newly generated numerical features for each value.

As shown in Figure 1, EMAC framework first learns primary value couplings to construct direct value coupling matrix $\mathbf{M}$ and indirect value coupling matrix $\mathbf{M}'$. Based on the pairwise primary value couplings, EMAC further employs the embedding method to obtain high-order complex value couplings through the value representation matrix $\mathbf{N}_v$. Distance notions like Euclidean distance can be utilised on the matrix $\mathbf{N}_v$. We expect to obtain an effective value representation matrix $\mathbf{N}_v$ that can well separate outlying values from others. Object outlierness can be subsequently measured based on matrix $\mathbf{N}_v$.

We propose EMAC framework to exhibit our insight of embedding high-order complex value couplings into value representation matrix. EMAC demonstrates good generalisability because multiple methods can be utilised to specify its components. EMAC also has potential applications to other machine learning tasks, e.g., data object representation. We introduce an instance of EMAC in the following section and verify its performance by empirical analysis.

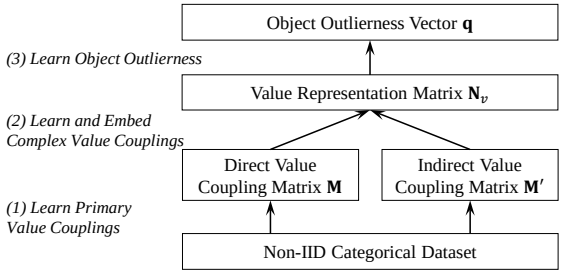

Figure 1: The EMAC Framework.

An EMAC Instance: SCAN

SCAN instantiates EMAC framework by specifying primary value coupling learning through Ochiai coefficient and conditional probability and specifying complex value coupling learning by defining value coupling bias and employing an extended Skip-gram architecture designed for network, i.e., *node2vec* (Grover and Leskovec 2016). We propose bidirectional selective value coupling learning to evaluate value outlierness and further measure object outlierness.

Learning Primary Value Couplings

SCAN learns primary value couplings from two aspects, i.e., direct and indirect couplings. Ochiai coefficient and conditional probability are popularly-used methods to measure pairwise direct value couplings in categorical data (Pang et al. 2017b; Xu et al. 2018a; 2018b). Ochiai coefficient-based matrix and conditional probability-based matrix $\mathbf{M}_1, \mathbf{M}_2 \in \mathbb{R}^{|\mathcal{V}| \times |\mathcal{V}|}$, are defined as follows.

$$\mathbf{M}_1(u, v) = \frac{P(u, v)}{\sqrt{P(u) \times P(v)}}, u, v \in \mathcal{V}, \quad (1)$$

$$\mathbf{M}_2(u, v) = \frac{P(u, v)}{P(u)}, u, v \in \mathcal{V}, \quad (2)$$

where $P(v)$ is the marginal probability of value v, i.e., $P(v) = |\{\boldsymbol{x} \in \mathcal{X} | v_f^{\boldsymbol{x}} = v\}|/N$, and $P(u, v)$ is the joint probability of value u and v, i.e., $P(u, v) = |\{\boldsymbol{x} \in \mathcal{X} | v_{f_u}^{\boldsymbol{x}} = u \cap v_{f_v}^{\boldsymbol{x}} = v\}|/N$.

On the other hand, indirect value coupling matrix $\mathbf{M}' \in \mathbb{R}^{|\mathcal{V}| \times |\mathcal{V}|}$ is calculated by the cosine similarity between conditional probability vectors.

$$\mathbf{M}'(u, v) = \frac{\mathbf{M}_2(u, \cdot) \cdot \mathbf{M}_2(v, \cdot)}{\|\mathbf{M}_2(u, \cdot)\| \|\mathbf{M}_2(v, \cdot)\|}, u, v \in \mathcal{V}, \quad (3)$$

where $\mathbf{M}_2(v, \cdot)$ denotes row vector of value v in matrix $\mathbf{M}_2$, and $\| \cdot \|$ is ℓ_2-Norm.

Learning Complex Value Couplings

In complex value coupling learning, we first calculate value coupling bias to enhance the capability of SCAN to better focus on outliers. Value coupling bias can enlarge the gap between outlying-to-outlying value couplings and others, which further facilitates complex value coupling learning to capture outlying values. Subsequently, a network embedding method is performed on the value network constructed by biased value couplings to represent each value by a newly generated numerical feature vector.

Value coupling bias is calculated via the integration of value clustering results and initial value outlierness. We use statistic results of clustering to acquire complex value couplings with different granularities which can demonstrate different semantics and well reflect the data characteristics. Initial value outlierness is scored through the bidirectional value coupling (BSVC)-based function.

Spectral clustering is performed on matrix $\mathbf{M}_2$ with setting different cluster number k, i.e., $\mathbf{D}_k = SC(\mathbf{M}_2, k)$, where $\mathbf{D}_k \in \mathbb{R}^{|\mathcal{V}|\times|\mathcal{V}|}$, $\mathbf{D}_k(u,v) = 1$ if value u and v are grouped into same cluster, and $\mathbf{D}_k(u,v) = 0$ if they are separated. In spectral clustering, we use default RBF kernel to obtain affinity matrix and apply discretisation to assign labels. Instead of setting a fixed value of k, we increase k from initial value 2 and stop increasing when appearing a cluster with only one member. Note that the dynamic setting of cluster number can capture different granularities of value couplings. Matrix $\mathbf{D} \in \mathbb{R}^{|\mathcal{V}|\times|\mathcal{V}|}$ is calculated based on the clustering results.

$$\mathbf{D} = \frac{1}{k_{max}-1}\sum_{i=2}^{k_{max}} \mathbf{D}_i, \tag{4}$$

where k_{max} is the cluster number when appearing a one-member cluster. Spectral clustering is chosen since it is efficient and is useful in hard non-convex clustering problems.

We first define BSVC-based value outlierness scoring function, and then introduce how to use it to calculate initial value outlierness.

Definition 1 (BSVC-based Value Scoring)**.** *BSVC-based value outlierness scoring function $\phi(\mathbf{C}, \mathcal{S}_o, \mathcal{S}_n, \alpha)$ is defined to get value outlierness vector $\eta \in \mathbb{R}^{|\mathcal{V}|}$.*

$$\eta = \phi(\mathbf{C}, \mathcal{S}_o, \mathcal{S}_n, \alpha) = \frac{1}{2\alpha|\mathcal{V}|}\Big(\sum_{v\in\mathcal{S}_o}\mathbf{C}(v,\cdot) + \Big(\mathbf{e} - \sum_{v\in\mathcal{S}_n}\mathbf{C}(v,\cdot)\Big)\Big), \tag{5}$$

where $\mathcal{S}_o$ and $\mathcal{S}_n$ are value subsets containing outlying values and normal values with size $\alpha|\mathcal{V}|$, $\mathbf{C} \in \mathbb{R}^{|\mathcal{V}|\times|\mathcal{V}|}$ is a value coupling matrix, and $\mathbf{e} = \sum_{i=1}^{|\mathcal{V}|} e_i$ is all-ones vector.

In order to get initial value outlierness vector η_0 through function ϕ, we rank values by rough value scoring function δ to obtain value subset $\mathcal{S}_o^\delta$ and $\mathcal{S}_n^\delta$, i.e, $\delta(v) = \frac{P(m)-P(v)}{P(m)}$, where m is the mode value of the same feature of v. Symmetric direct value coupling matrix $\mathbf{M}_1$ is employed as input matrix $\mathbf{C}$. α is set as a parameter of SCAN. Thus, initial value outlierness vector is obtained as $\eta_0 = \phi(\mathbf{M}_1, \mathcal{S}_o^\delta, \mathcal{S}_n^\delta, \alpha)$.

After getting clustering statistics matrix and value outlierness vector, we generate a non-zero value coupling bias matrix $\mathbf{B} \in \mathbb{R}^{|\mathcal{V}|\times|\mathcal{V}|}$ as follows.

$$\mathbf{B}(u,v) = (1 + \frac{\eta_0(u)+\eta_0(v)}{2}) \times (1 + \mathbf{D}(u,v)), u, v \in \mathcal{V}. \tag{6}$$

SCAN constructs an undirected weighted value network $G =< V, E >$ to further learn complex value couplings through network embedding. Each node in network is a feature value and edge weight represents the biased value couplings. The adjacency matrix of the network $\mathbf{A} \in \mathbb{R}^{|\mathcal{V}|\times|\mathcal{V}|}$ is defined as follows.

$$\mathbf{A} = \mathbf{M}_1 \circ \mathbf{M}' \circ \mathbf{B}, \tag{7}$$

where $\circ$ denotes entrywise product.

Various embedding methods are powerful to represent different kinds of data. In this scenario, network embedding can be employed to embed high-order complex value couplings into value representation matrix. Different kinds of relationships among the values in network G, which are originally represented by edges or other high-order topological measures of network, are captured by the distances between values in the newly generated vector space. The topological and structural characteristics of a value can also be encoded into its embedding vector. *node2vec* is performed on value network G to represent each value with a r-dimensional vector, i.e., generate a value representation matrix $\mathbf{N}_v \in \mathbb{R}^{|\mathcal{V}|\times r}$. Note that *node2vec* effectively preserves community structure as well as high-order proximity between nodes. Thus, it can yield value couplings between any two different values, i.e., the problem of capturing couplings of values from the same feature can also be addressed.

Evaluating Value and Object Outlierness

After getting value representation matrix $\mathbf{N}_v$, we propose Bidirectional Selective Value Coupling (BSVC) learning to evaluate value outlierness, which is shown in Figure 2. Final value outlierness vector η^* is subsequently used to obtain object outlier scores. BSVC learning iteratively performs BSVC-based value outlierness scoring and ranking-based value selection until finding a stationary value rank, and a stationary value outlierness vector is finally generated.

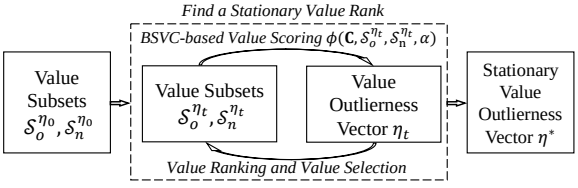

Figure 2: BSVC Learning to Evaluate Value Outlierness

The input value subsets $\mathcal{S}_o^{\eta_0}$ and $\mathcal{S}_n^{\eta_0}$ are the top α and bottom α values of the value rank sorted by initial value outlierness η_0. A value similarity matrix $\mathbf{M}_c \in \mathbb{R}^{|\mathcal{V}|\times|\mathcal{V}|}$ calculated from matrix $\mathbf{N}_v$, i.e., $\mathbf{M}_c(u,v) = \frac{\mathbf{N}_v(u,\cdot)\cdot\mathbf{N}_v(v,\cdot)}{\|\mathbf{N}_v(u,\cdot)\|\|\mathbf{N}_v(v,\cdot)\|}, u, v \in \mathcal{V}$, is set as the input coupling matrix of function ϕ. BSVC learning is denoted as $\eta^* = \Phi(\mathbf{M}_c, \mathcal{S}_o^{\eta_0}, \mathcal{S}_n^{\eta_0}, \alpha)$.

Note that BSVC learning is an optimisation of Selective Value Coupling learning framework (SelectVC) proposed in (Pang et al. 2017b). They are different in that: (i) BSVC learning evaluates value outlierness through value couplings from not only outlying values but normal values, while SelectVC only considers couplings from outlying values. (ii) BSVC learning stops iteration until the whole value rank is stationary, while SelectVC stops learning when the top-ranked outlying values being not changed. Thus, BSVC learning has fewer iteration times and severer convergence criterion to yield more precise value outlierness estimation.

After evaluating value outlierness, data object outlier score can be calculated through the summation of value outlierness, i.e., $\tau(\boldsymbol{x}) = \sum_{f \in \mathcal{F}} \eta^*(v_f^{\boldsymbol{x}}), \boldsymbol{x} \in \mathcal{X}$.

The Algorithm of SCAN

Algorithm 1 presents the procedure of SCAN. The input of SCAN is a set of data objects $\mathcal{X}$, subset size factor α, and representation dimensionality r. Step 1 learns direct and indirect primary value couplings. Steps 2-8 are performed to obtain value coupling bias matrix $\mathbf{B}$. *node2vec* is processed on network G on Step 10 to generate value representation matrix $\mathbf{N}_v$. Final value outlierness vector η^* and object outlierness vector τ are calculated through Steps 11-13. An object outlierness rank R is finally returned in Step 15.

Algorithm 1 *SCAN*

Input: $\mathcal{X}$ - data objects, α - subset size factor, r - representation dimensionality
Output: R - outlier rank

1: Generate $\mathbf{M}_1$, $\mathbf{M}_2$, and $\mathbf{M}'$ using Equation (1)(2)(3)
2: **repeat**
3: $\quad \mathbf{D}_k = SC(\mathbf{M}_2, k)$
4: $\quad k = k + 1$
5: **until** Appear a cluster with only one member
6: $\mathbf{D} \leftarrow \frac{1}{n} \sum_{i=2}^{n} \mathbf{D}_i$
7: $\eta_0 \leftarrow \phi(\mathbf{M}_1, \mathcal{S}_o^{\delta}, \mathcal{S}_n^{\delta}, \alpha)$
8: $\mathbf{B}(u,v) \leftarrow (1 + \frac{\eta_0(u)+\eta_0(v)}{2}) \times (1 + \mathbf{D}(u,v)), u, v \in \mathcal{V}$
9: Construct value network G as $\mathbf{A} \leftarrow \mathbf{M}_1 \circ \mathbf{M}' \circ \mathbf{B}$
10: Run *node2vec* on G to obtain matrix $\mathbf{N}_v \in \mathbb{R}^{|\mathcal{V}| \times r}$
11: $\mathbf{M}_c(u,v) \leftarrow \frac{\mathbf{N}_v(u,\cdot) \cdot \mathbf{N}_v(v,\cdot)}{\|\mathbf{N}_v(u,\cdot)\| \|\mathbf{N}_v(v,\cdot)\|}, u, v \in \mathcal{V}$
12: $\eta^* \leftarrow \Phi(\mathbf{M}_c, \mathcal{S}_o^{\eta_0}, \mathcal{S}_n^{\eta_0}, \alpha)$
13: $\tau(\boldsymbol{x}) \leftarrow \sum_{f \in \mathcal{F}} \eta^*(v_f^{\boldsymbol{x}}), \boldsymbol{x} \in \mathcal{X}$
14: $R \leftarrow$ Sort $\mathcal{X}$ w.r.t. τ in descending order
15: **return** R

Step 1 requires one scanning over the data objects, which has $O(|\mathcal{X}||\mathcal{V}|^2)$. Clustering process incurs the complexity of $O(k_{max}|\mathcal{V}|)$ in Steps 2-5. Constructing value network has $O(|\mathcal{V}||\mathcal{V}|)$. The time complexity of *node2vec* is $O(|\mathcal{V}|r)$. Steps 11-12 takes $O(|\mathcal{V}||\mathcal{V}|)$. The object scoring and sorting take $O(|\mathcal{X}||\mathcal{V}|)$ in Steps 13-14. Approximately, SCAN has linear time complexity w.r.t. number of data objects and is quadratic w.r.t. the number of features.

Experiments and Evaluation

Datasets

Thirteen publicly available real-world datasets are used, which cover diverse domains. Nine of these datasets are transformed from highly imbalanced data, where the smallest class is treated as outliers and the rest of classes or the largest class is normal. For the other four datasets, *Ada* and *MG* are transformed from balanced data by randomly sampling a small subset of the smallest class as outliers and keeping the largest class as normal class (imbalanced rate is controlled as 2%). The performance of these downsampled datasets is taken average over 10 times sampling; *StM* and *StP* are derived from a survey of math and Portuguese language courses in secondary school (Cortez and Silva 2008). In these two datasets, students with course grade less than 10 are treated as outliers, while students with course grade greater than 40 are normal objects (Xu et al. 2018a).

Performance Evaluation Methods

All the outlier detection methods in our experiments finally produce an object rank. Top-ranked data objects are the most likely outliers. Following (Pang et al. 2018b; Campos et al. 2016; Zimek et al. 2013), the quality of rank is evaluated by the area under ROC curve (AUC). AUC inherently takes the class-imbalance nature into consideration, making it comparable across different datasets in outlier detection (Campos et al. 2016). Higher AUC indicates better performance. The AUC would be around 0.5 given a random rank. We also employ the *Wilcoxon* signed rank test to examine the significance of AUC performance of SCAN against its competitors.

Following (Pang et al. 2017b; Jian et al. 2018; Xu et al. 2018a), three *data indicators*, i.e., average Mutual Information (MI), maximum feature efficiency (mfe) and average value relevance (avr), are defined to quantitatively measure the inherent characteristics of datasets, which are correlated with the performance of the outlier detectors.

MI is the average mutual information of all the features in a dataset. Mutual information measures how much knowing a feature reduces uncertainty about the other, which can be used to partially assess the non-IIDness of a dataset. Small MI indicates that the features demonstrate diversified information. mfe and avr are two indicators to measure the difficulties of a dataset. mfe is a feature-level indicator which is reported as the maximum AUC result of using frequency histogram of each feature to detect outliers. avr is a value-level indicator to evaluate the value correlation with outlier class label. All the values are sorted in descending order w.r.t. their conditional probability with outliers and the average conditional probability of top 20% values is calculated as avr. A dataset with low mfe and avr indicates that the dataset is very difficult.

Experiment Environment

All the experiments are executed at a 3.6GHz Desktop PC with 32GB memory. SCAN is implemented in Python. The source code of its competitors is obtained from their authors.

Table 1: A Summary of Datasets Used, Data Indicator Quantization Results and AUC Performance of SCAN and its Five Competitors. Data is ranked by indicator MI. The best performance for each dataset is boldfaced.

Data Infomation					Data Indicators			Outlier Detectors					
Data	Abbr.	$\lvert\mathcal{F}\rvert$	$\lvert\mathcal{V}\rvert$	$\lvert\mathcal{O}\rvert$	MI	mfe	avr	SCAN	HOUR	CBRW	POP	LeSiNN	iForest
SylvaP	Syl	87	174	457	0.0018	0.7889	0.3182	**0.9885 ± 0.0036**	0.9721	0.9689	0.7635	0.9557	0.8748
Celeba	Cele	39	78	202599	0.0182	0.7961	0.0888	**0.9005 ± 0.0066**	0.8879	0.8462	0.8968	0.7777	0.7015
Student Portuguese	StP	30	124	457	0.0246	0.8248	0.0829	**0.9406 ± 0.0160**	0.8405	0.8151	0.9167	0.8849	0.8532
Caltech16	Cal	253	506	829	0.0299	0.9780	0.3980	**0.9952 ± 0.0006**	0.9933	0.9925	0.9928	0.9903	0.9705
Solar	Sol	12	42	1066	0.0353	0.8221	0.2688	**0.8852 ± 0.0087**	0.5324	0.8812	0.8527	0.8534	0.8403
Student Math	StM	30	127	242	0.0397	0.6824	0.1522	**0.7358 ± 0.0363**	0.6072	0.4996	0.6584	0.5834	0.5980
BreastC	BrC	9	41	286	0.0678	0.6545	0.6138	**0.7292 ± 0.0077**	0.6867	0.6064	0.4726	0.6741	0.6440
aYahoo	aY	33	66	450	0.1055	0.9673	1.0000	**0.9993 ± 0.0001**	0.9678	0.9988	0.9988	0.9890	0.9902
Seismic	Sei	10	73	2584	0.1146	0.7470	0.4148	**0.7532 ± 0.0022**	0.7479	0.7350	0.7297	0.7272	0.7238
mammographic	MG	4	19	568	0.1316	0.8021	0.3899	**0.8857 ± 0.0037**	0.8785	0.8667	0.8489	0.7366	0.7387
adaP	Ada	9	112	3773	0.1888	0.6119	0.3330	**0.7298 ± 0.0604**	0.6055	0.5879	0.4089	0.4793	0.4992
Credit	Cre	9	77	30000	0.3045	0.6969	0.7852	0.7092 ± 0.0090	**0.7202**	0.5804	0.4109	0.6717	0.6396
BreastW	BrW	9	89	683	0.4829	0.9358	1.0000	**0.9939± 0.0001**	0.9898	0.9918	0.9907	0.9851	0.9754
							Average	**0.8651 ± 0.0119**	0.8023	0.7977	0.7647	0.7930	0.7730
							p-value	-	0.0024	0.0002	0.0002	0.0002	0.0002

HOUR, CBRW, and POP are in JAVA. iForest and LeSiNN are in MATLAB.[1]

Effectiveness of SCAN in Real-world Data

Experiment Settings. SCAN is compared with five state-of-the-art outlier detection methods, i.e., HOUR (Pang et al. 2017a), CBRW (Pang, Cao, and Chen 2016), POP (Pang et al. 2017b), LeSiNN (Pang, Ting, and Albrecht 2015), and iForest (Liu, Ting, and Zhou 2012) on the thirteen real-world datasets to examine its effectiveness[2]. SCAN uses $\alpha = 0.15$ and $r = 128$ by default. Its contenders use recommended parameter settings. HOUR, CBRW, and POP are closely related methods. They also model feature value couplings to detect outliers in categorical data, and they are chosen to examine the significance of complex value coupling learning of SCAN. LeSiNN is based on least similar nearest neighbours, and iForest is a feature subspace-based method. Note that both of them work on numerical data. We use one-hot transformation method to convert categorical features to binary features to allow them to process on the same datasets (Xu et al. 2018a; Campos et al. 2016). SCAN, LeSiNN and iForest are non-deterministic methods (i.e., their performance may have slight difference between two runs), hence we report the average AUC result over 10 independent runs.

Findings: SCAN Significantly Outperforming Five State-of-the-art Outlier Detectors. The AUC performance of SCAN and its five competitors is reported in Table 1. The standard deviation of SCAN is also reported. Our method SCAN achieves the best performance on twelve datasets and significantly outperforms its five contenders at the 99% confidence level. Averagely, our method obtains 8%, 8%, 13%, 9%, and 12% AUC improvement over HOUR, CBRW, POP, LeSiNN, and iForest, respectively.

The superiority of SCAN is mainly because it successfully model, represent and utilise high-order complex value couplings to better capture the non-IIDness nature of data, which substantially improves its performance on the datasets with low MI, e.g., *StP* and *StM*. Note that other three coupling-based methods HOUR, CBRW and POP may obtain comparably good performance with SCAN and outperforms LesiNN and iForest on some datasets, e.g., *Cele* and *MG*. However, their performance is suboptimal and unstable because they only focus on the pairwise primary value couplings and fails to capture reliable value relationships that hide in high-order complex value couplings. In particular, SCAN impressively obtains over 66%, 47% and 54% AUC improvement over HOUR, CBRW and POP on *Solar*, *StM* and *BrC*, respectively. It is interesting that all the outlier detectors can perform very well on datasets *Syl*, *Cal*, *aY*, and *BrW*. This may be due to high mfe and/or avr of these datasets, i.e., they have at least one highly relevant feature and/or a small group of values that can well indicate outliers. On the contrary, it is very challenging for outlier detectors to obtain good performance on datasets with low mfe and/or avr, e.g., *Sei* and *Ada*. Nevertheless, SCAN also exhibits its superiority compared with existing methods on these complex datasets.

Resilience of SCAN to Noise

Experiment Settings. Following (Zimek, Schubert, and Kriegel 2012; Pang et al. 2018a), we generate a group of synthetic data (90 normal data objects and 10 outliers described by 20 features) to examine the resilience of SCAN to datasets with different level of noise. Relevant features are binary. They contain a normal value with frequency 90 and a relevant value with frequency 10 that has fixed conditional

[1]Our experiments show that POP runs comparably fast than LeSiNN and iForest, and runs one to two orders of magnitude faster than CBRW and HOUR. Thus, We reimplement POP in Python and compare the efficiency of SCAN and POP. Our method SCAN runs approximately one to two orders of magnitude slower than POP. Due to space limitations, the detailed efficiency results are omitted.

[2]Two classical outlier detection algorithms, i.e., CompreX (Akoglu et al. 2012) and LSA (He, Deng, and Xu 2005), are also performed. The results show that SCAN significantly outperform them at the 95% confidence level. Also, the empirical results in (Pang, Cao, and Chen 2016; Pang et al. 2017a) show that CBRW and HOUR significantly outperforms them. Therefore, we focus on the comparison with latest and closely-related outlier detectors.

probability with outliers (0.5 is set in the experiments). All the data objects obey Gaussian distributions in other irrelevant features. Equal width discretisation is employed to convert these numerical features to categorical features. All the six outlier detectors are examined on the datasets with different percentage of relevant features. In order to have more reliable results, we generate 10 datasets for each noise level and the average AUC over them is reported.

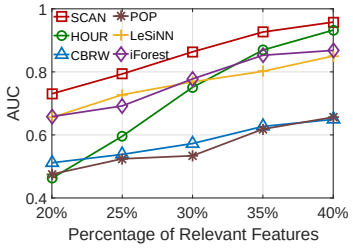

Figure 3: AUC Performance of SCAN and its Competitors on Datasets with Different Percentage of Relevant Features.

Findings: SCAN has Good Resilience to Noise. The AUC performance of SCAN and its five competitors on the synthetic datasets are shown in Figure 3. The AUC results of SCAN are very close to 1 when the percentage of relevant features is more than 40%. All the outlier detectors cannot obtain effective results when the datasets contain less than 20% relevant features, i.e., their AUC performance is around 0.5. Therefore, we focus on the noisy datasets with 20%-40% relevant features. SCAN performs consistently better than HOUR, CBRW, POP, LeSiNN, and iForest on these datasets, which further validates the effectiveness of modeling complex value couplings to alleviate the interference from noise. Note that HOUR is specially designed for noisy data, it gradually exhibits its superiority with the increasing of percentage of relevant features, and obtains comparably good result with SCAN on dataset with 40% relevant features. However, it performs very poorly on the datasets with overwhelming noise level. The performance of CBRW and POP are considerably downgraded by the noise, since CBRW is a full space-based method and POP only models pairwise value couplings through conditional probability. It is interesting that LeSiNN and iForest can obtain quite good performance on these datasets. It is may because LeSiNN and iForest are based on subsampling, and they are less sensitive to noise. However, they are still markedly outperformed by SCAN.

Sensitivity Test

Experiment Settings: We examine the sensitivity of SCAN w.r.t. its parameters r and α on all the datasets.

Findings: SCAN Performing Stably w.r.t. its Parameters. The AUC results and standard deviation (SD) of SCAN with varying α and r are reported in Figure 4. SCAN shows stable performance on all the datasets. Here we selectively demonstrate the AUC performance w.r.t. α and r on five datasets, i.e., *Stp*, *Cele*, *Sol*, *aY*, and *BrW*, due to the space limits. Note that the parameters can be tuned based on the specific prior knowledge when the SCAN is applied in different domains. Parameter α is related to multiple factors, e.g., the outlier proportion, noise rate, and the number of outlying values contained per outlier (Pang et al. 2017b). $2\alpha|\mathcal{V}|$ is the subspace size that is used in the value outlierness evaluation of SCAN. Too large α normally makes value subspace lose its meaning. As for Parameter r, it partially determines the efficiency of SCAN. In general, $\alpha = 0.15$ and $r = 128$ are recommended in practice.

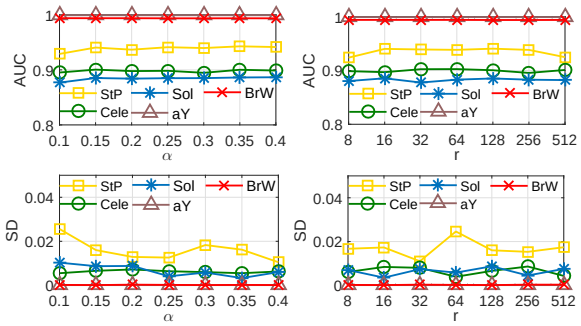

Figure 4: Sensitivity Test Results of SCAN w.r.t. α and r.

Conclusions

This paper introduces EMAC framework to propose an insight of using embedding method to learn complex value couplings for detecting outliers in non-IID categorical data. We further propose SCAN as an instance of EMAC. SCAN first models primary value couplings, and then defines coupling bias to capture complex couplings with different granularities. High-order complex value couplings can be further learnt and embedded in the value representation matrix by performing the network embedding method on the biased value coupling-based value network. In addition, a value subspace-based value outlierness evaluation method (i.e., BSVC learning) is proposed to show how to utilise obtained value couplings to detect outliers. Our extensive experiments show that SCAN significantly outperforms five state-of-the-art outlier detectors on thirteen real-world datasets, and has much better resilience to datasets with different level of noise. Besides, SCAN performs stably w.r.t. its parameters. In future, in order to practically use SCAN to detect outliers in multi-source data, we plan to implement SCAN in distributed system JointCloud (Wang, Shi, and Zhang 2017).

Acknowledgements

This work is supported by the National Key Research and Development Program of China (2016YFB1000101), and by the National Natural Science Foundation of China (No.61472439 and No.61379052).

References

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

Aggarwal, C. C. 2017. *Outlier analysis*. Springer.

Akoglu, L.; Tong, H.; Vreeken, J.; and Faloutsos, C. 2012. Fast and reliable anomaly detection in categorical data. In *CIKM*, 415–424. ACM.

Campos, G. O.; Zimek, A.; Sander, J.; Campello, R. J.; Micenková, B.; Schubert, E.; Assent, I.; and Houle, M. E. 2016. On the evaluation of unsupervised outlier detection: measures, datasets, and an empirical study. *Data Mining and Knowledge Discovery* 30(4):891–927.

Cao, L., and Yu, P. S. 2016. Non-iid recommendation theories and systems. *IEEE Intelligent Systems* 31(2):81–4.

Cao, L.; Ou, Y.; and Yu, P. S. 2012. Coupled behavior analysis with applications. *IEEE Transactions on Knowledge and Data Engineering* 24(8):1378–1392.

Cao, L. 2014. Non-iidness learning in behavioral and social data. *The Computer Journal* 57(9):1358–1370.

Chen, T.; Tang, L.-A.; Sun, Y.; Chen, Z.; and Zhang, K. 2016. Entity embedding-based anomaly detection for heterogeneous categorical events. In *IJCAI*, 1396–1403. AAAI Press.

Cinbis, R. G.; Verbeek, J.; and Schmid, C. 2016. Approximate fisher kernels of non-iid image models for image categorization. *IEEE transactions on pattern analysis and machine intelligence* 38(6):1084–1098.

Cortez, P., and Silva, A. M. G. 2008. Using data mining to predict secondary school student performance. In *FUBUTEC*, 5–12. EUROSIS.

Das, K.; Schneider, J.; and Neill, D. B. 2008. Anomaly pattern detection in categorical datasets. In *SIGKDD*, 169–176. ACM.

Grover, A., and Leskovec, J. 2016. node2vec: Scalable feature learning for networks. In *SIGKDD*, 855–864. ACM.

He, Z.; Xu, X.; Huang, Z. J.; and Deng, S. 2005. FP-outlier: Frequent pattern based outlier detection. *Computer Science and Information Systems* 2(1):103–118.

He, Z.; Deng, S.; and Xu, X. 2005. An optimization model for outlier detection in categorical data. In *Advances in Intelligent Computing*. Springer. 400–409.

Jian, S.; Pang, G.; Cao, L.; Lu, K.; and Gao, H. 2018. Cure: Flexible categorical data representation by hierarchical coupling learning. *IEEE Transactions on Knowledge and Data Engineering*.

Keller, F.; Müller, E.; and Bohm, K. 2012. HiCS: High contrast subspaces for density-based outlier ranking. In *ICDE*, 1037–1048. IEEE.

Lazarevic, A., and Kumar, V. 2005. Feature bagging for outlier detection. In *SIGKDD*, 157–166. ACM.

Liu, F. T.; Ting, K. M.; and Zhou, Z.-H. 2012. Isolation-based anomaly detection. *ACM Transactions on Knowledge Discovery from Data* 6(1):1–39.

Pang, G.; Cao, L.; Chen, L.; and Liu, H. 2017a. Learning homophily couplings from non-iid data for joint feature selection and noise-resilient outlier detection. In *IJCAI*, 2585–2591. AAAI Press.

Pang, G.; Xu, H.; Cao, L.; and Zhao, W. 2017b. Selective value coupling learning for detecting outliers in high-dimensional categorical data. In *CIKM*, 807–816. ACM.

Pang, G.; Cao, L.; Chen, L.; Lian, D.; and Liu, H. 2018a. Sparse modeling-based sequential ensemble learning for effective outlier detection in high-dimensional numeric data. In *AAAI*, 3892–3899.

Pang, G.; Cao, L.; Chen, L.; and Liu, H. 2018b. Learning representations of ultrahigh-dimensional data for random distance-based outlier detection. In *SIGKDD*, 2041–2050. ACM.

Pang, G.; Cao, L.; and Chen, L. 2016. Outlier detection in complex categorical data by modelling the feature value couplings. In *IJCAI*, 1902–1908. AAAI Press.

Pang, G.; Ting, K. M.; and Albrecht, D. 2015. LeSiNN: Detecting anomalies by identifying least similar nearest neighbours. In *ICDM Workshop*, 623–630. IEEE.

Sathe, S., and Aggarwal, C. C. 2016. Subspace outlier detection in linear time with randomized hashing. In *ICDM*, 459–468. IEEE.

Wang, H.; Shi, P.; and Zhang, Y. 2017. Jointcloud: A cross-cloud cooperation architecture for integrated internet service customization. In *ICDCS*, 1846–1855. IEEE.

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

Xu, H.; Wang, Y.; Cheng, L.; Wang, Y.; and Ma, X. 2018a. Exploring a high-quality outlying feature value set for noise-resilient outlier detection in categorical data. In *CIKM*, 17–26. ACM.

Xu, H.; Wang, Y.; Wu, Z.; Ma, X.; and Qin, Z. 2018b. Combine value clustering and weighted value coupling learning for outlier detection in categorical data. In *DEXA*, 439–449. Springer.

Zhao, W.; Li, Q.; Zhu, C.; Song, J.; Liu, X.; and Yin, J. 2018. Model-aware categorical data embedding: a data-driven approach. *Soft Computing* 22(11):3603–3619.

Zhou, Z.-H.; Sun, Y.-Y.; and Li, Y.-F. 2009. Multi-instance learning by treating instances as non-iid samples. In *Proceedings of the 26th annual international conference on machine learning*, 1249–1256. ACM.

Zhu, C.; Cao, L.; Liu, Q.; Yin, J.; and Kumar, V. 2018. Heterogeneous metric learning of categorical data with hierarchical couplings. *IEEE Transactions on Knowledge and Data Engineering* 30(7):1254–1267.

Zimek, A.; Gaudet, M.; Campello, R. J.; and Sander, J. 2013. Subsampling for efficient and effective unsupervised outlier detection ensembles. In *SIGKDD*, 428–436. ACM.

Zimek, A.; Schubert, E.; and Kriegel, H.-P. 2012. A survey on unsupervised outlier detection in high-dimensional numerical data. *Statistical Analysis and Data Mining* 5(5):363–387.