Anomaly Detection Methods for Categorical Data: A Review

AYMAN TAHA, Faculty of Computers and Information, Cairo University, Egypt
ALI S. HADI, American University in Cairo, Egypt, and Cornell University, Ithaca, NY, USA

Anomaly detection has numerous applications in diverse fields. For example, it has been widely used for discovering network intrusions and malicious events. It has also been used in numerous other applications such as identifying medical malpractice or credit fraud. Detection of anomalies in quantitative data has received a considerable attention in the literature and has a venerable history. By contrast, and despite the widespread availability use of categorical data in practice, anomaly detection in categorical data has received relatively little attention as compared to quantitative data. This is because detection of anomalies in categorical data is a challenging problem. Some anomaly detection techniques depend on identifying a representative pattern then measuring distances between objects and this pattern. Objects that are far from this pattern are declared as anomalies. However, identifying patterns and measuring distances are not easy in categorical data compared with quantitative data. Fortunately, several papers focussing on the detection of anomalies in categorical data have been published in the recent literature. In this article, we provide a comprehensive review of the research on the anomaly detection problem in categorical data. Previous review articles focus on either the statistics literature or the machine learning and computer science literature. This review article combines both literatures. We review 36 methods for the detection of anomalies in categorical data in both literatures and classify them into 12 different categories based on the conceptual definition of anomalies they use. For each approach, we survey anomaly detection methods, and then show the similarities and differences among them. We emphasize two important issues, the number of parameters each method requires and its time complexity. The first issue is critical, because the performance of these methods are sensitive to the choice of these parameters. The time complexity is also very important in real applications especially in big data applications. We report the time complexity if it is reported by the authors of the methods. If it is not, then we derive it ourselves and report it in this article. In addition, we discuss the common problems and the future directions of the anomaly detection in categorical data.

CCS Concepts: • **Information systems applications → Data mining**;

Additional Key Words and Phrases: Computational complexity, data mining, holo entropy, intrusion detection systems, mixed data, outliers detection, nominal data, novelty detection, semi-supervised learning, Shannon entropy, supervised learning, unsupervised learning

ACM Reference format:
Ayman Taha and Ali S. Hadi. 2019. Anomaly Detection Methods for Categorical Data: A Review. *ACM Comput. Surv.* 52, 2, Article 38 (May 2019), 35 pages.
https://doi.org/10.1145/3312739

Authors' addresses: A. Taha, Faculty of Computers and Information, Cairo University, 5 Dr Ahmed Zewail St., Orman, Giza 12613, Egypt; email: a.taha@fci-cu.edu.eg; Ali S. Hadi, The American University in Cairo, P.O.Box 74, New Cairo 11835, Egypt; email: ahadi@aucegypt.edu or ali-hadi@cornell.edu.
ACM acknowledges that this contribution was authored or co-authored by an employee, contractor or affiliate of a national government. As such, the Government retains a nonexclusive, royalty-free right to publish or reproduce this article, or to allow others to do so, for Government purposes only.
© 2019 Association for Computing Machinery.
0360-0300/2019/05-ART38 $15.00
https://doi.org/10.1145/3312739

1 INTRODUCTION

Anomalies are a minority of objects (observations, cases, or points) that are inconsistent with the pattern suggested by the majority of objects in the same dataset. It is very important to identify anomalies in data, because they can distort the analysis and the decisions based on the analysis [23, 30, 43, 81–83, 85]. Anomaly detection (or outlier or novelty detection as it is sometimes referred to in the literature) is a well-studied problem in different research areas, such as communications, statistics, data mining, and machine learning (see, e.g., References [16, 33, 45, 80, 97, 115, 151]).

The massive increases in e-commerce applications, cloud computing services, and remote business access have triggered the need for protection systems against unauthorized access. In communications, illegal access to network resources is called intrusion. Intrusion Detection Systems (IDS) are substantial parts in network security that observe computer networks to detect odd activities [53]. IDS are classified into two main categories: Signature-based and Anomaly-based.

A signature-based intrusion detection system relies on building a database of defined signatures for known attacks. It raises an alert when observing signature similar to those attacks in its database. It fails to detect new attacks whose signatures do not exist in the database. However, an anomaly-based intrusion detection system usually builds a statistical model defining normal activities. Abnormal activities (anomalies) are those activities that are discordant with the defined model [181].

Real datasets often consist of different types of variables, that is, some variables are quantitative and others are qualitative or categorical. These data are called mixed data. The analysis of quantitative data has a venerable history and, by comparison, categorical data has received less attention than quantitative data. Categorical data, however, are common in many different domains (e.g., biomedical, educational, sociological, psychological, political, and social sciences [7, 8]. Anomaly detection techniques for categorical data make use of data within these fields [15, 172].

There are two types of categorical data: nominal and ordinal. Examples of the former type include gender, nationality, and type of network protocol. Examples of the latter type include letter grade in a course (e.g., A, B, C, D, F), network traffic volume (e.g., low, medium, high), and a Likert scale variable (e.g., 1 = strongly disagree, 2 = disagree, 3 = neutral, 4 = agree, and 5 = strongly agree). Finding anomalies in ordinal variables is relatively easier than the identification of anomalies in nominal data, because the categories in nominal variables have no natural ordering [174].

1.1 Applications of Anomaly Detection in Categorical Data

Applications of anomaly detection in categorical data are numerous. We mention here just a few examples:

- Network Intrusions: Detecting users who have abnormal accesses is an interesting application for anomaly detection in cloud computing [154]. Here users' profiles and behaviors are usually coded as categorical variables.
- Moving Objects: Anomaly detection has useful applications in analyzing moving objects data such as identifying objects that have unexpected moves and identifying types of road segments, where moving objects have anomalous behavior [71, 116].
- Medical and Health Data: Anomaly detection methods provide useful information to decision makers such as the identification of excessively high medical payments and/or medical malpractice. This information can be used to improve health care management [96]).
- Social Networks: There are many applications that make use of anomaly detection in social networks such as finding anomalous users in social groups who have different viewpoints, interests, and believes [5, 69].

- Credit Fraud: Unauthorized credit card use may be identified when an unexpected behavior is detected. Categorical attributes are often used to represent behavioral data [175].
- Questionnaire Data: Responses to questionnaires are usually mapped into categorical or ordinal attributes. The importance of anomaly detection in questionnaire data is highlighted in Reference [187].
- Earth Science: Finding anomalies in spatiotemporal data, e.g., weather patterns or climate changes in various geographical areas give an explanation behind interesting spatiotemporal patterns [117].
- Law Enforcement: Examples of anomaly detection applications for law enforcement are discovering anomalies in trading activities and insurance claims [3].

1.2 Challenges Facing Anomaly Detection in Categorical Data

The identification of anomalies in categorical data faces some challenges. These include:

(1) Anomaly detection methods focus on identifying the pattern (statistical distribution) suggested by the majority of observations, then considering the observations that do not follow the assumed pattern as anomalies [30]. Some distance functions are proposed in the literature to compute distance between categorical observations (e.g., References [32, 41]). Identifying a pattern and measuring a distance are not easy in categorical data. Consequently, anomaly detection methods are more common for quantitative than for categorical data.
(2) Several alternative but different definitions of anomalies in categorical data exist in the literature [179]. Anomaly detection methods can identify different sets of observations as anomalies depending on the definition they adopt.
(3) The number of benchmark datasets that can be used to test the performance of anomaly detection methods for categorical data (e.g., computation time, detection rate, etc.) is very small. In addition, it is not straightforward to generate synthetic categorical data with known anomalies because of the lack of methods that generate such data [171].
(4) Computational complexity is also a challenging problem in the identification of anomalies especially in categorical data, because most real applications have huge datasets in terms of the number of observations, the number of categorical variables, and the number of categories in each. Consequently, time complexity is a significant issue when applying anomaly detection methods to categorical data.

1.3 Related Work

The problem of anomaly detection has been reviewed in a number of survey articles, books and book chapters (see, e.g., References [38, 39, 46, 80, 91, 100, 122, 123]). In a recent book [3], one can find an account of various outliers detection approaches for quantitative as well as categorical data. Chapter 8 of Reference [3], in particular, discusses the problem of identifying outliers in categorical data.

Some survey articles are generic, giving a broad overview of anomaly detection methods (see, e.g., References [59, 60, 72, 119, 163]). Other review articles focus only on one type of methods based on application, technical approach and/or data type. These methods can be classified according to application domain as follows:

- Network Intrusion Detection: These articles survey, compare, classify and/or identify challenges in intrusion detection methods in computer networks (see, e.g., References [10, 20, 61, 70, 75, 78, 108, 134]). Distance and similarity measures used in intrusion detection methods are surveyed in Reference [177].

- Wireless Sensor Network (WSN): This group of articles focus on anomaly detection techniques for wireless sensor data (see, e.g., References [1, 2, 49, 118, 121, 145, 146, 149, 161, 185]). Characteristics of anomaly detection techniques for wireless sensor networks in a non-stationary and harsh environments are surveyed in References [130, 159]. Anomaly detection in smart home and automated surveillance are reviewed in References [21, 165].
- Data Streams: Few review articles study the features and classification of anomaly detection in data stream (see, e.g., References [19, 55, 66]). Anomaly detection methods for dynamic streaming data are reviewed in Reference [184]. Also, Reference [141] focuses on anomalies in continuous time variant data stream.
- Fraud Detection: Fraud and abuse detection methods have useful applications in many domains. Health care fraud detection techniques are surveyed in References [48, 99]. Fraud detection methods for biological and chemical data are reviewed in Reference [46]. Fault detection methods in industrial processes are studied in Reference [27].
- Financial, Business and Recommender Systems: Financial and credit card fraud detection methods are surveyed in References [11, 101, 135, 138, 178]. Outlying profile attacks in recommender systems are studied in Reference [58].

Another set of survey articles focus on the approach or the methodologies that are used to identify outliers. These approaches include:

- Data Mining Approach: There are several review articles that survey data mining-based anomaly detection methods (see, e.g., References [12, 26, 61, 99, 102]). Other review articles cover a certain data mining pattern (such as neural network-based novelty detection methods are covered in Reference [123]), while, frequent pattern-based anomaly detection methods are reviewed in Reference [18]. Clustering-based anomaly detection methods are reviewed in Reference [55]. Anomaly detection techniques for distributed data are studied in Reference [12]. Frequent pattern-based anomaly detection methods and their scoring measures are discussed in Reference [153].
- Machine Learning Approach: There are few survey articles that cover machine learning-based anomaly detection methods (see, e.g., References [20, 27, 139]).
- Statistical Approach: The anomaly detection problem appeared in the statistical literature before data mining and machine learning. Statistical anomaly detection methods concentrate on simple data types (such as numerical and quantitative data) rather than complex structures (such as categorical, graphical and/or spatial data). Statistical anomaly detection methods are surveyed in References [46, 52, 72, 84, 122].

A third group of review articles concentrate on the type of data where anomaly detection methods are proposed to identify outliers within this type of data. These articles can be classified as follows:

- Social Networks and Graph Data: Anomaly detection methods in social media are compared and reviewed in some survey articles (see, e.g., References [102, 148, 156, 183]). However, some surveys focus on wider scope, graph data (see, e.g., References [14, 143, 155]).
- Time Series and Spatial Data: Several anomaly detection methods are proposed for time series and spatial data (see, e.g., References [19, 79, 80, 141]).
- Complex and Big Data: Big and complex data are active areas of research particularly in the computer science literature. Recently, several methods are proposed for the identification of anomalies in high dimensional, complex, and big data (see, e.g., References [57, 65]). Anomaly detection for high-dimensional data are given in References [4, 188]. Other techniques

focus on the identification of anomalies in subspaces rather than studying all dimensions [166].

- Empirical Comparative Reviews: Some articles focus on experimental comparative study among a certain type of anomaly detection methods such as some statistical and distance-based anomaly detection methods (see, e.g., References [22, 132]). Further, empirical evaluations of unsupervised anomaly detection measures are conducted in Reference [36].

Most of the survey articles for anomaly detection provide extensive surveys for detection of anomalies in quantitative data (see, e.g., Reference [9, 38, 42, 84, 91, 100]). Some of these surveys mention categorical data but just in passing. Discrete sequence anomaly detection methods are reviewed in Reference [39].

1.4 Contributions

This article provides a comprehensive review of the research on anomaly detection in categorical data. Particular contributions are:

(1) The interest in the identification of anomalies in categorical data is recent, but fortunately it has been intensifying during the last decade. To the best of our knowledge, this is the first review article that focusses on the identification of anomalies in categorical data.
(2) The papers on the identification of anomalies in categorical data appear in both the statistical literature as well as the computer science literature such as machine learning and big data. This article reviews both of these literatures.
(3) The existing anomaly detection methods for categorical data are categorized into different approaches. Each anomaly detection method is reviewed and discussed. Discussions of the number of parameters they require and comparisons among these methods are provided in Sections 2 to 10.
(4) Since computational complexity is a significant issue in real applications, for each method, we report the time complexity if it is reported by the author(s) of the methods. If it is not, then we derive and report it in this article. Tables 1 and 2 show the reviewed anomaly detection methods together with their computational complexities and required parameters.
(5) Extensive discussion of the problems facing anomaly detection methods in categorical data are provided and topics for future research in anomaly detection in categorical data are presented.

1.5 Organization

The rest of this article is organized as follows: Sections 2–10 give a review of anomaly detection methods in categorical data. Discussions are provided in Section 11. Section 12 provides concluding remarks and recommendations for future work. Tables 1 and 2 summarize the reviewed anomaly detection methods for categorical data, their complexities, and their required parameters.

2 METHODS BASED ON INDICATOR VARIABLES

One approach for the detection of anomalies in categorical data is to represent categorical data by numeric values. Then anomaly detection for quantitative data can be used. Examples of methods that follow this approach are (a) Indicator Variables (IV) [162] and (b) a method based on Multiple Correspondence Analysis (MCA) [162].

The IV method replaces each categorical variable by indicator variables [162]. For example, a categorical variable X_j with c_j categories, can be represented by c_j indicator (binary) variables, $Y_{n\times c_j} = \{Y_1^j, Y_2^j, \ldots, Y_{c_j}^j\}$, where Y_r^j is the rth indicator variable corresponding to the rth category

Table 1. Various Anomaly Detection Algorithms and their Complexities

Approach	Method	Complexity
Indicator Variables	IV (•)	$O(nq + n^2 \sum_{j=1}^{q}(c_j - 1))$
	MCA	$O(nq + 2n^2Q + nQ^2 + Q^3)$
Marginal Frequency	AVF	$O(nq)$
	SCF	$O(nq)$
	WAVF	$\approx O(nq)$
	WDOD	$\approx O(nq)$
	CMBOD	$\approx O(nq)$
	Bouguessa's method (*) (•)	$\approx O(nq + \text{Max}_m(nkqi + im(t_E + t_M)))$
Itemset Frequency	FPOF (•)	$\approx O(n(2^Q))$
	LOADED (*) (•)	$\approx O(n(2^Q))$
	ODMAD (*) (•)	$\approx O(n(2^Q))$
	FNDI-OD (•)	$\approx O(n(2^Q))$
Diversified Frequency	CBRW-OD (•)	$O(nQ^2 + \lvert E\rvert I_{max} + nq)$
Bayesian/Conditional Frequecy	CA (•)	$\approx O(n(2^Q) + (q^w))$
	APD (•)	$\approx O(n(2^Q) + (q^w))$
	RHH (*) (•)	$\approx O(n(2^Q) + (q^w))$
	AA	$\approx O(n((2^Q)) + (C_2^{\lvert f\rvert}) + (\lvert R\rvert)^2)$
Density-Based Methods	HOT	$\approx O(nq^2 + \lvert H\rvert q\ max(\lvert he\rvert))$
	k-LOF (*)	$O(n^2(q + \sum_{i=1}^{k} i))$
	WATCH (*)	$O(ngs + q^2gt/2)$
Clustering-Based Methods	ROAD (•)	$O(nqc_{max} + nqk^2 + nqkt + n\ log\ n)$
	Rough-ROAD (•)	$O(nqc_{max} + nqk^2 + nqkt + n\ log\ n)$
Distance-Based Methods	Orca (*)	$O(qn^2)$ linear in practice
	CNB	$O(n^2(k + S(\theta) + q) + n(k + M))$
	RBRP	$O(nq \log n + qn^2)$
	iOrca	$O(qn^2)$
Shannon Entropy	LSA (•)	$O(nqC_M^n)$
	GA	$O(nqM)$
Holo Entropy	ITB-SP	$O(nq)$
	EEB-SP	$O(nq)$
	EEB-SS	$O(Mnq)$
	EEB-SS	$O(Mnq)$
Compression	KRIMP	$\simeq O(n(2^Q + f^2(nQ + 1))))$
	CompreX	$\approx O(nq^2)$
SemiSupervised	FraC (*)	$\approx O(n^2q)$
	SAnDCat	$\simeq O(nq^2 \log q + n^2 + n \log n)$

(*) This complexity is for the categorical part of methods for mixed data.
(•) This complexity is derived and reported in this article.

of the jth categorical variable X_j. To identify the anomalies in this binary dataset with Q indicator variables, where Q is the total number of categories in the dataset, the Canberra distance [64] is used. For each observation x_i in the binary dataset, it computes an outlying score, S_i, as the average of the Canberra distance between x_i and all other $n - 1$ observations in the dataset. It declares an observation x_i as an anomaly if $S_i > \theta$, where θ is a predefined parameter chosen by the user [162].

Table 2. Various Anomaly Detection Algorithms and their Required Parameters

Approach	Algorithm	Reference	No. of Param.	Decision Parameters	Intrinsic Parameters
Indicator Variables	**IV**	[162]	1	θ	
	MCA	[162]	1	θ	
Marginal Frequency	**AVF**	[106]	1	M	
	SCF	[173]	1	M	
	WAVF	[150]	1	M	
	WDOD	[186]	1	θ	
	CMBOD	[109]	1	M	
	Bouguessa's(∗)	[34]	3	$M_{\max}$, knn	knn
Itemset Frequency	**FPOF**	[89]	3	M	*minfreq, maxlen*
	LOADED(∗)	[131]	4	win, $\triangle S$	*minfreq, maxlen*
	ODMAD(∗)	[105]	4	win, $\triangle S$	*minfreq, maxlen*
	FNDI-OD	[107]	3	M	*minfreq, maxlen*
Diversified Frequency	CBRW-OD	[133]	2	M	α
Bayesian/Conditional Frequecy	**CA**	[50]	5	θ	*minfreq, maxlen*, α, β_α
	APD	[51]	5	θ	*minfreq, maxlen*, α, β_α
	RHH(∗)	[144]	4	θ	*minfreq, maxlen*, α
	AA	[127]	4	θ	*minfreq, maxlen, minconf*
Density-based	**HOT**	[176]	3	θ	*minfreq, maxlen*
	k-LOF (*)	[182]	2	θ	k-walks
	WATCH	[112]	1	M	
Clustering-based	**ROAD**	[169]	3	M	α, K
	Rough-ROAD	[170]	5	M	ε, w_l, K, α
Distance-based	**Orca(∗)**	[24]	2	M	knn
	CNB	[113]	3	M	knn, sim
	RBRP	[74]	3	M	knn, K
	iOrca	[28]	2	M	knn
Shannon Entropy	**LSA**	[87]	1	M	
	GA	[88]	1	M	
Holo Entropy	**ITB-SP**	[180]	1	M	
	EEB-SP	[152]	1	M	
	ITB-SS	[180]	1	M	
	EEB-SS	[152]	1	M	
Compression	**KRIMP**	[164]	3	θ	*minfreq, maxlen*
	CompreX	[15]	1	M	
SemiSupervised	**FraC(∗)**	[129]	1	θ	
	SAnDCat	[95]	2	θ	k

(*) The parameters required for the categorical part of methods for mixed data.

The MCA is a generalization of Correspondence Analysis (CA) when the number of categorical variables is greater than two [25]. The MCA-based anomaly detection method represents categorical variables by indicator variables [162] and use the Canberra distance to identify the anomalies in the transformed numerical datasets. The MCA steps are then applied to the binary matrix B of size $n \times \sum_{j=1}^{q} c_j$ instead of the contingency table, which is used by the CA.

The time complexity of the Indicator Variables method can be divided into two parts. The time complexity of creating the indicator variables, which is $O(nq)$, and the time complexity

of computing the Canberra distance, which is $O(n^2 \sum_{j=1}^{q}(c_j - 1))$. Accordingly, the total time complexity for the Indicator Variable method is $O(nq + n^2 \sum_{j=1}^{q}(c_j - 1))$.

The time complexity of the MCA-based method consists of three parts. First, the time complexity of creating the indicator variables is $O(nq)$. Second, the time complexity of finding the singular value decomposition is $(nQ^2 + n^2Q + Q^3)$ [54, 76]. Finally, the time complexity of computing the Canberra distance is $O(n^2Q)$. Therefore, the total time complexity is $O(nq + 2n^2Q + nQ^2 + Q^3)$.

As can be seen, the Indicator Variable method is computationally expensive due to the substantial increase of number of indicator variables and the quadratic cost of measuring distance with respect to the number of observations. The use of indicator variables method is not recommended when the number of categories is large. The authors solve this problem by grouping similar categories into one category using some clustering methods. However, this solution slightly decreases the number of indicator variables but it leads to loss of information due to grouping [162].

The MCA-based anomaly detection method is also computationally expensive due to the computation of singular value decomposition in addition to the quadratic cost of measuring distance with respect to the number of observations.

3 FREQUENCY-BASED METHODS

Frequency-based methods make use of the frequency (i.e., the number of occurrences of categories) instead of distances. There are three types of frequencies that can be used to identify anomalies in categorical data: (a) Marginal frequency, (b) Itemset Frequency, and (c) Diversified Frequency.

3.1 Marginal Frequency

Anomalies in categorical data can be defined as observations with small marginal frequency. Methods using this definition are (a) the Attribute Value Frequency (AVF) [106] and [105], (b) the Square of the Complement Frequency (SCF) [173], (c) the Weighted Attribute Value Frequency (WAVF) [150], (d) the Weighted Density-based Outlier Detection (WDOD) method [186], (e) a Cloud Model-based Outlier Detection method (CMBOD) [109], and (f) Bouguessa's method [34].

The AVF is based on computing a frequency score for each observation, x_i, as

$$\mathrm{AVF}(x_i) = \frac{1}{q}\sum_{j=1}^{q} f(x_{ij}), \tag{1}$$

where $f(x_{ij})$ is the marginal frequency of x_{ij} within the variable X_j. The AVF declares the M objects with the lowest AVF scores as outliers, where M is a parameter chosen by the user. Alternatively, Reference [147] suggest estimating the number of anomalies M by assuming that the frequency scores follow a normal distribution.

The Square of the Complement Frequency (SCF) calculates an outlying score for each observation, x_i, in the dataset as

$$\mathrm{SCF}(x_i) = \sum_{j=1}^{q} \frac{[1 - p(x_{ij})]^2}{c_j}, \tag{2}$$

where c_j is the number of categories in the variable X_j and $p(x_{ij})$ is the marginal relative frequency of x_{ij} (the number of occurrence within X_j divided by number of observations). It then declares the M observations with the highest outlying scores as outliers. However, SCF uses the square of the complement frequency to increase the difference between frequent and infrequent categories. Moreover, SCF takes into consideration the number of categories (c_j) to give variables with small c_j higher weights in the score function.

The AVF and SCF methods do not take into account the sparseness of the frequencies in categorical variables. Therefore the performance of these methods can be improved by capturing the sparsity of categorical variables by giving weight to the frequencies. Two examples of such weight are (a) the Weighted Attribute Value Frequency (WAVF) [150] and (b) the Weighted Density-based Outlier Detection (WDOD) method [186].

In the WAVF a weighting function is presented to indicate the variable sparsity in measuring the outlying scores. The higher sparsity the variable has, the higher impact on the outlying score. Thus observations with more sparse categories have higher probability for being outliers. The sparseness level of a categorical variable can be measured by sparsity statistical functions, e.g., the range or the standard deviation of the marginal frequencies. The Weighted Attribute Value Frequency (WAVF) is proposed to extend the AVF method as

$$\mathrm{WAVF}(x_i) = \frac{1}{q}\sum_{j=1}^{q} f(x_{ij}) * R_j, \tag{3}$$

where R_j is the range of the frequencies of the jth categorical variable.

Another way of giving weights to the frequencies is the weighted density-based outlier detection (WDOD) method for categorical data [186]. It relies on estimating the density of each categorical variable then it computes the weighted density for the whole dataset. More specifically, WDOD starts with computing a weight, $W(X_j)$, for each variable X_j as

$$W(X_j) = \frac{1 - E^c(X_j)}{\sum_{l=1}^{q}(1 - E^c(X_l))}, \tag{4}$$

where

$$E^c(X_j) = \sum_{i=1}^{c_j} p(x_{ij}) * (1 - p(x_{ij})), \tag{5}$$

is the complement entropy for X_j, $p(x_{ij})$ is the relative frequency of a category x_{ij}, and c_j is number of categories in X_j. According to Reference [186], the complement entropy, $E^c(X_j)$, is proposed in [114] to measure the information gain or uncertainty to give different weights to categorical variables to emphasize their importance. Unlike the logarithmic behavior of the Shannon entropy, the complement entropy measures both the uncertainty and the fuzziness. The WDOD score for an object x_i is then obtained by the sum of the weighted relative frequencies, that is,

$$\mathrm{WDOD}(x_i) = \sum_{j=1}^{q} p(x_{ij}) * W(X_j). \tag{6}$$

Accordingly, $\mathrm{WDOD}(x_i)$ takes into consideration the uncertainty in the density of each variable. The lower weighted density $\mathrm{WDOD}(x_i)$ an object has, the higher probability of being a density-based outlier. Therefore, the WDOD declares an observation, x_i, as an outlier if $(\mathrm{WDOD}(x_i) < \theta)$, where θ is a predefined parameter.

The Cloud Model-based Outlier Detection method (CMBOD) [109] is based on building a cloud model that summarizes a dataset into a multidimensional cloud model [110]. The method consists of three steps. Step 1 transforms categorical variables to numerical values through a probabilistic mapping method called *cloud drops*. For a categorical variable X_j with c_j categories, each category is replaced by its marginal relative frequency $p(x_{ij})$.

Step 2 uses the transformed dataset in building multidimensional cloud model as follows: Three digital characteristics of the dataset are extracted. The digital characteristics of a variable X_j are (a) the average of X_j, $\bar{x}_j$, (b) the entropy of X_j, E_{nj}, and (c) the hyper entropy of X_j, E_{hj} [111]. The

entropy of an observation is calculated as $E_{nj}(x_{ij}) = \sqrt{\frac{\pi}{2}} * |(x_{ij} - \bar{x}_j)|$. The entropy of a variable X_j is the average entropies of all observations, which is

$$E_{nj} = \frac{1}{n}\sum_{i=1}^{n} E_{nj}(x_{ij}), \quad j = 1, 2, \ldots, q. \tag{7}$$

The hyper entropy, E_{hj}, of a dataset is the average measure for the fussiness of entropy values. It is calculated as $E_{hj} = \sqrt{S_j^2 - E_{nj}^2}$ [86], where

$$S_j^2 = \frac{1}{n-1}\sum_{i=1}^{n}(x_{ij} - \bar{x}_j)^2, \quad j = 1, 2, \ldots, q \tag{8}$$

is the variance of X_j.

Step 3 computes the certainty degree for each observation x_i as

$$cd(x_i) = \exp\left(\sum_{j=1}^{q} \frac{-(x_{ij} - \bar{x}_j)^2}{2E_{nj}^2}\right), \quad i = 1, 2, \ldots, n. \tag{9}$$

The certainty degree $cd(x_i)$ is used as a measure of belonging to the established cloud model. The lower certainty degree means higher deviation from the established cloud model, that is, the higher probability of being an anomaly. The CMBOD labels the M observations with the smallest $cd(x_i)$ as outliers.

The method proposed in Reference [34] is designed to detect anomalies in mixed datasets. It is based on probabilistic approach that employs univariate beta mixture model (the bivariate beta mixture model in the case of mixed data) to distinguish outlying from non-outlying data. The method computes two outlying scores for each observation, one score for the quantitative attributes and the other score for the categorical attributes.

The outlying score of the p quantitative (numerical) variables is

$$S_N(x_i) = \sum_{j=1}^{p} \log\left(1 + \sum_{r=1}^{k}[x_{ij} - \mathrm{knn}_r(x_{ij})]^2\right), \tag{10}$$

where p is number of quantitative (numerical) variables and $\mathrm{knn}_r(x_{ij})$ is the value of the rth nearest neighbor to (x_{ij}) in the jth quantitative variable. Whereas the outlying score of the q categorical part is

$$S_C(x_i) = \sum_{j=1}^{q} \log(f(x_{ij})), \tag{11}$$

where q is the number of categorical variables and $f(x_{ij})$ is the frequency of (x_{ij}) in the jth categorical variable. The outlying scores use the log to emphasize the difference between frequent and rare values. Since large values of $S_N(x_i)$ indicates outliers but small values of $S_C(x_i)$ indicate outliers. Therefore, $S_C(x_i)$ is reversed and replaced by

$$S_C(x_i) \leftarrow \max(S_C(x_i)) - S_C(x_i), \tag{12}$$

so that large values of $S_C(x_i)$ now indicate outliers. Each of the two scores are then normalized to have values between 0 and 1. A bivariate beta distribution can then be used to determine if the normalized scores are large enough to be declared as outliers.

The author claims that this method does not require any decision parameters but this claim is not entirely true, since it requires three decision parameters. It is true that the author avoids requiring the number of outliers, M, by using a mixture of m univariate beta components. Then, it

labels observations that corresponds to the vectors with the highest scores as outliers. It estimates number of components, m, by trying each possible values of m from $1, 2, \ldots, M_{\max}$, which is a user predefined parameter. In addition, in each iteration, it partitions the scores by the K-means method then it applies Expectation Maximization (EM) Algorithm to estimate the parameters of the m components. Thus, the second parameter is K in the K-means method. A third parameter is k, which is required for the k nearest-neighbors (knn) method, but the author suggests using $k = \sqrt{n}$ as a default value.

The time complexity of the AVF, SCF, WAVF, and WDOD is $\approx O(nq)$, which linearly increases with the number of observations and the number of categorical variables.

The time complexity of the CMBOD method can be computed as follows: In each step, it scans the dataset once and it has three steps (three scans). Consequently, the time complexity of CMBOD method is $\approx O(nq)$, which linearly increases with the number of observations and the number of categorical variables.

The time complexity of Bouguessa's method is composed of two parts: Computing outlying scores and identification of anomalies. It requires $O(nq)$ in computing outlying scores. However, the complexity of the K-means method by using Lloyd's Algorithm [120] is $O(nkqi)$, where i is number of iterations. While, the complexity of EM algorithm relies on the number of iterations and time for processing the expectation and maximization steps [29]. So, it approximately requires $O(im(t_E + t_M))$, where m is number of parameters to be estimated, t_E is the time required to do the expectation step, and t_M is the time required for the maximization step of the EM algorithm. Consequently, the total complexity of Bouguessa's method is $\approx O(nq + \text{Max}_m(nkqi + im(t_E + t_M)))$. Accordingly, it requires complex time processing. Another serious disadvantage of Bouguessa's method is that its output (identified outliers) is random, that is, the identified outliers can change from one run to the next due to the random nature of the K-means method.

The AVF, SCF, WAVF, WDOD, and CMBOD are fast and scalable methods (they can efficiently deal with large scale categorical datasets), but they only consider the marginal frequency and ignore any dependency structure among the categorical variables. Moreover, they require specifying the number of anomalies M in advance, which is impractical in real applications. In addition to the difficulty of defining the suitable value of these parameters, the results are very sensitive to these parameters.

3.2 Itemset Frequency

The methods in this category consider the frequency of itemsets (combinations of categories of size $\leq$ *maxlen*, a predefined parameter). The anomalies are defined as observations that have small itemset frequencies. These methods start with building a set of frequent itemsets (itemsets with frequency $\geq$ a certain parameter *minfreq*), then declare observations with less frequent itemsets as outliers. Examples of these methods are: (a) Frequent Pattern Outlier Factor (FPOF) [89], (b) Link-based Outlier and Anomaly Detection in Evolving Datasets (LOADED) [73, 131], (c) Outlier Detection for Mixed Attribute Datasets (ODMAD) [105], and (d) Frequent Non-Derivable Itemsets—Outlier Detection (FNDI-OD) [107].

The FPOF, the categorical part of LOADED, and the categorical part of ODMAD define anomalies as observations that contain infrequent patterns in their itemsets. For defining the list of frequent itemsets (S), they use two parameters: *minfreq* (minimum frequency of frequent itemset) and *maxlen* (maximum number of categories in the itemset). In addition to itemsets parameters, they use another parameter for declaring anomalies (e.g., FPOF declares M observations with infrequent itemsets as outliers; LOADED and ODMAD require two extra parameters; the window size (win) and the maximum difference score ($\triangle S$). For each observation x_i, they define the window of x_i (a set of observations of size win). They compute the average outlying score in its

window, then calculate the difference between the average of window scores and x_i's score. If the difference exceeds $\triangle S$, then they declare x_i as an outlier.

A significant problem with the itemset-based methods is that the number of frequent itemsets tends to be huge. For each observation, the set of frequent items is scanned to identify frequent items belonging to this observation. Thus, they take a very long time in processing. To reduce the running time of itemset-based anomaly detection methods, FNDI-OD replaces the set of all frequent itemsets with a condensed and compacted version called Non-Derivable Itemsets (NDI) by removing frequent derivable itemsets that can be deduced with their frequency from non-derivable itemsets. NDI represents the minimal frequent itemsets cover that can be used to reproduce all frequent itemsets. For example, if we have three frequent itemsets a, ab, and abc with the same frequency 10, then ab is a derivable itemset, because its frequency can be derived from a and abc, while a and abc are non-derivable. The size of NDI is usually very small compared to the size of its relevant set of frequent items. FNDI-OD scans NDI for each observation. As a result, it significantly reduces the processing time compared to other itemset-based anomaly detection methods. At the same time, the FNDI-OD has similar accuracy to the FPOF [107].

The time-complexity of itemset frequency-based methods can be divided into two parts; Finding frequent itemsets and calculating outlying scores. An *a priori* method is used for finding the frequent itemsets [6]. Its time complexity is

$$O\Big(n\Big(C_1^Q + C_2^{Q-u} + C_3^{Q-u} + \cdots + C_{maxlen}^{Q-u}\Big)\Big) \approx O(n(2^{(Q-u)})), \tag{13}$$

where Q is the total number of categories and u is the number of infrequent categories [62]. In the worst case, all categories are frequent, then $u = 0$. Accordingly, the complexity of the Apriori method in the worst case is $\approx O(n(2^Q))$. The time complexity of calculating outlying scores is $O(n|S|)$. The total time complexity for these methods is then $O(n(2^Q + |S|))$. The FNDI-OD reduces the processing time of itemset-based anomaly detection methods, but in the worst case it has the same time complexity.

The itemset frequency-based methods take into account dependency among categorical variables, since they handle categorical variables jointly by considering itemset frequencies instead of only marginal frequencies. However, they have the following disadvantages: They are computationally expensive as they have exponential complexity with respect to the number of categorical variables and linear complexity with respect to the size of the set of frequent itemsets, which grows combinatorially. Moreover, they require two parameters; *maxlen* and *minfreq*. Their results are very sensitive to the values of these parameters.

3.3 Diversified Frequency

Under this category, we discuss one method for identifying anomalies in categorical data, that is, Couple Biased Random Walk (CBRW) [133]. The CBRW method takes into account intra-feature coupling (distribution) among categories at the same categorical variable as well as inter-feature coupling (interactions) among categories at different categorical variables. First, to begin with, let m_j be the modal (the most frequent) category for the categorical variable X_j and $p(m_j)$ be the relative frequency of the modal category of the categorical variable X_j. Then, for each category x_{ij} in X_j, we compute an intra-feature coupling deviation score as

$$\delta(x_{ij}) = [dev(x_{ij}) + base(m_j)]/2, \tag{14}$$

where $dev(x_{ij}) = [p(m_j) - p(x_{ij})]/p(m_j)$ and $base(m_j) = 1 - p(m_j)$. Thus, the intra-feature coupling considers both the marginal frequencies of categories as well as the whole frequency distribution within each variable.

Second, a directed graph, $\mathcal{G}$, is built to capture the inter-feature value coupling (interactions) among categorical values. In this graph, each category is represented as a node, while inter-feature coupling is propagated from a node u to another node v through an edge connecting u and v,

$$A(u,v) = p(u|v) = \frac{p(u,v)}{p(v)}. \tag{15}$$

The inter-feature coupling checks wether a certain category coupled with other outlying categories from other categorical variables. $A(u,v)$ measures the strength of coupling between u and v. Finally, based on $\delta(v)$ is as defined in Equation (14), and $A(u,v)$, CBRW builds a biased random walk matrix W_b as

$$W_b(u,v) = \frac{\delta(v)A(u,v)}{\sum_{v \in V} \delta(v)A(u,v)}, \tag{16}$$

where V is the set of nodes in $\mathcal{G}$ (all categories in the whole dataset). $W_b(u,v)$ represents the transition from the node u to the node v having a probability proportional to $\delta(v)A(u,v)$. Thus, each random walk is biased by the value of $\delta(v)$.

After building W_b, CBRW generates the probability distribution of biased random walk column vector, π_0, to calculate an outlying score for each category. Then CBRW initializes π_0 values by a uniform distribution. Thereafter, it computes π_{t+1} at time $t+1$ on the basis of π_t as

$$\pi_{t+1} = (1-\alpha)\frac{1}{|V|}1 + \alpha\, W_b^t \pi_t, \tag{17}$$

where α is a damping factor to guarantee convergence. It is shown in Reference [133] that if $\mathcal{G}$ is irreducible and aperiodic, then π converges to a unique stationary probability, π^*, regardless of the initialization values π_0. Subsequently, the outlying score for a certain category v is computed as

$$\text{CBRW}(v) = \pi^*(v). \tag{18}$$

The computed outlying CBRW scores can be used in features selection for high dimensional data and/or identifying outliers. CBRW scores can measure the importance of a certain categorical variable as

$$Rel(X_j) = \sum_{i=1}^{c_j} \text{CBRW}(x_{ij}), \tag{19}$$

where (x_{ij}) is the ith category in the jth variable and c_j is number of categories in the jth variable. Thus, it can be utilized in ranking and selection of the most relevant categorical variables then it could be embedded into any outliers detection method.

Moreover, CBRW scores can identify outliers by computing an outlying score for each observations (x_i) as

$$\text{CBRW-OD}(x_i) = \sum_{j=1}^{q} W_j * \text{CBRW}(x_{ij}), \tag{20}$$

where

$$W_j = \frac{Rel(X_j)}{\sum_{j=1}^{q} Rel(X_j)}. \tag{21}$$

Accordingly, CBRW labels observations having the highest M CBRW-OD scores as outliers.

The time complexity of the CBRW method is composed of three parts. First, building the biased random walk matrix W_b requires $O(nQ^2)$, *where* Q is total number of categories in all variables. Second, the complexity of computing the probability distribution of biased random walk vector is $O(|E|I_{max})$, where $|E|$ is number of edges in $\mathcal{G}$ and I_{max} is maximum number of iterations

for calculating π^*. Finally, the complexity of finding the outliers is $O(nq)$. Therefore, the overall complexity of CBRW is $O(nQ^2 + |E|I_{max} + nq)$.

One advantage of the CBRW is that it takes into consideration correlation among the categorical variables, since it captures frequency distribution of categorical variables as well as inter-feature coupling. In addition, it ranks categorical variables according to its relevance to outlierness. However, it is computationally expensive as it has a quadratic complexity with respect to the number of categories. Moreover, it requires two parameters: the number of outliers M and the damping factor α.

4 BAYESIAN/CONDITIONAL FREQUENCY-BASED METHODS

This approach has a different definition of categorical anomalies. It searches for observations with high marginal frequencies and small joint frequencies. In other words, it defines anomalies to be observations with infrequent combination of categories while its categories are frequent. The conditional anomalies are quite different from anomalies identified by other anomaly detection methods. They look for observations that contain frequent categories that are rarely observed together. Examples of methods that follow this approach are (a) Conditional Algorithm (CA) [50], (b) the Anomaly Pattern Detection (APD), (c) the method proposed in [144] (denoted here by RHH, the first letters of the authors' names), and (d) the Attribute Association (AA) algorithm [127].

The Conditional Algorithm (CA) defines outliers as observations with infrequent combinations of frequent categories [50]. The CA calculates a measure of rareness, named $r(x_{ij}, x_{ik})$ ratio, between a pair of categories x_{ij} and x_{ik} in the categorical variables X_j and X_k, respectively. The $r(x_{ij}, x_{ik})$ ratio is defined as

$$r(x_{ij}, x_{ik}) = \frac{p(x_{ij}, x_{ik})}{p(x_{ij})p(x_{ik})}, \tag{22}$$

where $p(x_{ij})$ and $p(x_{ij}, x_{ik})$ are the marginal and joint relative frequencies, respectively. A small r-ratio means that the multiplication of marginal probabilities is much higher than their joint probability, which indicates a higher probability of anomalous co-occurrence. The r-ratio is applied to itemsets, a subset of categorical attributes, up to *maxlen* length. For each observation, r-ratio is computed for every pair of itemsets. Then CA assigns the smallest r-ratio to be its outlying score, which means the outlying score reflects a subset of variables rather than considering all variables.

The CA method labels observations that have outlying scores less than a certain threshold θ as anomalies. The number of itemsets is huge, because it grows combinatorially with the number of categorical variables, q. Therefore, the conditional algorithm tries to decrease the number of candidate subsets by pruning itemsets from weakly correlated variables. It employs the mutual information [158]

$$\mu(X_j, X_l) = -\sum_{i=1}^{i=c_j}\sum_{k=1}^{k=c_l} p(X_j = x_{ij} \wedge X_l = x_{kl}) \times \log\left(\frac{p(X_j = x_{ij} \wedge X_l = x_{kl})}{p(X_j = x_{ij}) \times p(X_l = x_{kl})}\right), \tag{23}$$

which measures the correlation between a pair of categorical variables X_j and X_l. It only considers itemsets within strongly correlated variables whose mutual information, μ, is greater than predefined threshold, β_μ. Consequently, it dramatically decreases the number of r-ratio computations.

The Anomaly Pattern Detection (APD) method is proposed to detect anomalous patterns, groups of related observations having outliers percentage higher than expected [51]. The APD method uses the CA algorithm [50] as first step to identify individual outliers. Then it uses rule-based technique to study the behavior of outliers in each pattern.

Similar to the conditional probability method, the RHH method [144] searches for observations with frequent attribute values but infrequent joint co-occurrence. It can find anomalies in

categorical datasets as well as mixed dataset. In the mixed datasets, it transforms quantitative variables into categorical variables by discretizing quantitative variables into fixed length intervals. In the training phase, it builds Bayesian network (see, e.g., References [37, 126]) to capture the dependency among the attributes. It takes advantage of the parent-child relationship of Bayesian network and stores the counts of categories and combinations of categories using the Alternating Decision (AD) Tree [125]. It computes the same outlying score as the conditional algorithm.

The Attribute Association (AA) algorithm [127] defines conditional anomalies as observations that contain frequent categories but these itemsets are rarely observed together. The method starts with deriving a set of association rules with high confidence from the data. Then it computes an outlying score called outlier degree as

$$od(x_i) = \frac{|C^+_{x_i} - C^0_{x_i}|}{|C^0_{x_i}|}, \tag{24}$$

where $C^0_{x_i}$ and $C^+_{x_i}$ are initial and final cover sets for x_i. For details about how to construct these cover sets, see Reference [127]. The method labels an observation x_i as an outlier if $od(x_i) > \theta$, where θ is a predefined threshold.

The time complexity of the CA, the RHH, and the outlier detection part of the APD can be divided into two parts: Finding frequent itemsets and calculating the r-ratio by building the Bayesian network. The time complexity for finding frequent itemsets is $O(n(2^Q))$, where Q is the total number of categories. However, the complexity of building Bayesian network is $\approx O(q^w)$, where w is the tree width [47]. Therefore, the total complexity of the conditional algorithm and the RHH method is $\approx O(n(2^Q) + (q^w))$.

However, the time complexity for the AA algorithm consists of three parts: Finding frequent itemsets, identifying a set of association rules with high confidence, and building the final cover. The first and most expensive part is finding frequent itemsets. The time complexity for finding frequent itemsets is $\approx O(n((2^Q))$. The complexity of identifying confident association rules is $\binom{|f|}{2}$, where $|f|$ is the length of frequent itemsets, which is combinatorial in the number of frequent itemsets f. The time complexity for finding the final cover is $(|R|)^2$, where $|R|$ is the length of the set of high confidence association rules. Therefore, the total complexity of the AA method is $\approx O(n((2^Q)) + (C_2^{|f|}) + (|R|)^2)$.

Conditional anomalies detection methods have time complexity problems, since they require a combinatorial time for building itemsets and a long time for searching in the high conditional probability space. Moreover, they require many parameters (see, e.g., *minfreq*, *maxlen*, and *minconf*).

5 DENSITY-BASED METHODS

Density-based anomalies or local anomalies approach aims at identifying observations that have outlying behavior in local areas, in which observations usually share similar characteristics [35]. Local anomalies are different from global anomalies, which are inconsistent with the pattern suggested by the majority of all other observations, not only observations within their local areas [35, 44, 98, 182]. Local anomaly detection methods for categorical data include (a) the Hyperedge-based Outlier Test (HOT) [176], (b) the k-Local Anomalies Factor k-LOF [182], and (c) the WATCH method [112].

The HOT method is based on defining two sets of variables: Common variables, C, and Outlying variables, A. HOT ignores other variables. It develops a graph called hypergraph that captures the similarity among categorical observations. The hypergraph is divided into groups (named hyperedges, H) based on the frequent itemsets in C. Each hyperedge $he \in H$ contains a set of observations

that has the corresponding itemset he. Then, for each observation x_i in the hyperedge (he) and an attribute A_j in a set of outlying attributes A, HOT calculates a deviation measure,

$$Dev^{he}(x_i, A_j) = \frac{S^{he}_{A_j}(c_r) - \mu_{S^{he}_{A_j}}}{\sigma_{S^{he}_{A_j}}}, \tag{25}$$

where c_r is the value of x_i in the jth categorical variable and $S^{he}_{A_j}(x_i)$ is the number of observations in hyperedge he, which have the value c_r in the categorical attribute A_j,

$$\mu_{S^{he}_{A_j}} = \frac{1}{|A^{he}_j|} \sum_{c_r \in A^{he}_j} S^{he}_{A_j}(c_r) \tag{26}$$

is the mean value of $S^{he}_{A_j}(c_r)$, and

$$\sigma_{S^{he}_{A_j}} = \sqrt{\frac{1}{|A^{he}_j|} \sum_{c_r \in A^{he}_j} \left(S^{he}_{A_j} - \mu_{S^{he}_{A_j}}\right)^2} \tag{27}$$

is the standard deviation of $S^{he}_{A_j}(c_r)$. HOT declares an observation x_i as an outlier for the hyperedge he and outlying attributes A_j if $Dev^{he}(x_i, A_j) < \theta$, where θ is a predefined deviation parameter.

The k-LOF is a local anomaly detection method for both categorical and quantitative data [182]. It extends Local Anomalies Factor (LOF) [35] to categorical domain. The k-LOF identifies an observation as a local outlier if its relationships with its neighbors are weaker than the relationships between its neighbors and its neighbors' neighbors. It measures the similarity among observations in the categorical data by a weighted undirected graph named similarity graph, $G = (V, E, w)$, where V is a set of vertices (observations) in the graph G, E is a set of edges (unordered pairs of vertices), and w is the weight of similarity between two vertices.

The k-LOF uses the concept of k-walk, which is the connection of length k edges between two observations. $N^k(x_i)$ is the set of neighbors connected to x_i within k-walks including x_i. The k-LOF defines the similarity of k-walk between two observations x_i and x_j, as

$$s^k(x_i, x_j) = \begin{cases} w(x_i, x_j), & \text{if } (x_i, x_j) \in E \text{ and } k = 1, \\ \sum\limits_{(x_m, x_j) \in E} w(x_m, x_j) \times s^{k-1}(x_i, x_m), & \text{if } k > 1. \end{cases} \tag{28}$$

In addition, the k-LOF computes another type of similarity named the accumulated similarity of k-walk between two observations x_i and x_j, as

$$S^k(x_i, x_j) = \sum_{i=1}^{k} s^i(x_i, x_j). \tag{29}$$

Then, the above two similarity measures are combined in an outlying score for x_i, which is given by

$$k\text{-LOF}(x_i) = \frac{1}{S^k(x_i, x_i) \times n} \sum_{j=0}^{n} S^k(x_i, x_j). \tag{30}$$

The k-LOF labels an observation, x_i, as an outlier if $k\text{-LOF}(x_i) > \theta$, where θ is a predefined parameter. The k-LOF takes into consideration the direct relationships between an observation and its direct neighbors as well as the indirect relationships among the neighbors and the neighbors' neighbors where $k > 1$. Thus, it requires the parameters θ and the maximum length of indirect relationships k.

Most recently, the WATCH method for identifying outliers in high dimensional categorical datasets using feature grouping is proposed in Reference [112]. The method aims at detecting local outliers (outliers in subspaces of the full-dimensional space). WATCH consists of two phases: Feature grouping and outliers detection. In the feature grouping phase, it partitions the set of q categorical variables into g disjoint feature groups by collecting correlated variables into the same group. Let

$$\mathrm{FR}(X_j, X_l) = \frac{\mathcal{I}(X_j, X_l)}{\mathcal{E}(X_j, X_l)} \tag{31}$$

be the feature relation between two features X_j and X_l, where

$$\mathcal{I}(X_j, X_l) = -\sum_{i=1}^{i=c_j}\sum_{k=1}^{k=c_l} p(X_j = x_{ij} \wedge X_l = x_{kl}) \times \log\left(\frac{p(X_j = x_{ij} \wedge X_l = x_{kl})}{p(X_j = x_{ij}) \times p(X_l = x_{kl})}\right), \tag{32}$$

is the Mutual Information [158], which measures the correlation between a pair of categorical variables X_j and X_l, and

$$\mathcal{E}(X_j, X_l) = -\sum_{i=1}^{i=c_j}\sum_{k=1}^{k=c_l} p(X_j = x_{ij} \wedge X_l = x_{kl}) \times \log(p(X_j = x_{ij} \wedge X_l = x_{kl})) \tag{33}$$

is the Shannon Entropy [160] between a pair of categorical variables X_j and X_l.

WATCH performs the following steps, which are similar to the K-means clustering algorithm [67], to cluster the categorical variables into $g \in 2, 3, \ldots, q/2$ feature groups: Initialization of the pivot dimensions, allocating each feature to one of the feature groups, and updating the selection of the pivot dimensions. During the initialization of the pivot dimensions, a random feature is selected to be the pivot dimension, η_1, for the first feature group, $\mathcal{G}_1$, then the next pivot dimension $\eta_i, i \in 2, 3, \ldots, g$, is the feature that has the minimum sum of feature relations with all the selected pivot dimensions, $\sum_{k=1}^{i-1} \mathrm{FR}(\eta_i, \eta_k)$.

In the allocation step, for each categorical variable, X_j, WATCH searches for the optimal feature group $\mathcal{G}_i$ that has the maximum feature relation with its pivot dimension, $\mathrm{FR}(\eta_i, X_j)$. Then, it updates the choice of the pivot dimensions in each feature group according to the allocation step. Finally, it iterates the allocation of each categorical attribute to the optimal (best) feature group and updating the pivot dimensions until there is no change in the pivot dimensions and hence, there is no change in feature groups.

For any specific partition of $g \in 2, 3, \ldots, q/2$ groups, the aggregate relation,

$$\mathrm{AR(D)} = \sum_{i=1}^{i=g}\sum_{r \in \mathcal{G}_i} \mathrm{FR}(\eta_i, X_r), \tag{34}$$

is computed for the whole data set. WATCH chooses the value of g that maximizes the aggregate relation, that is,

$$g = \mathrm{argmax}_{g \in 2,3,\ldots,q/2} \mathrm{AR}(D). \tag{35}$$

The second phase of WATCH method is detecting outliers in each feature group. It calculates a weighting factor, $w(X_j)$ for each categorical variable according to the feature correlation between this variable and other variables within its feature group,

$$w(X_j) = \frac{1}{r}\sum_{l=1}^{r} \mathrm{FR}(X_l, X_j), \tag{36}$$

where r is the number of categorical variables in the feature group and $\mathrm{FR}(X_l, X_j)$ is the feature relation between X_l and X_j. Then, it computes an outlying score for each observation according to each feature group $\mathcal{G}_k$ as

$$\mathrm{WATCH}_{\mathcal{G}_k}(x_i) = \frac{1}{r}\sum_{j=1}^{r}\begin{cases} 0, & \text{if } f(x_{ij}) = 1, \\ w(X_j) \times \log\left(\frac{(f(x_{ij})-1)^{(f(x_{ij})-1)}}{f(x_{ij})^{f(x_{ij})}}\right), & \text{otherwise.} \end{cases} \tag{37}$$

The outlying score, $\mathrm{WATCH}_{\mathcal{G}_k}(x_i)$, takes negative values. Furthermore, the higher outlying score an observation has, the higher probability of being an outlier. Thus, WATCH labels observations having the highest M scores as outliers with respect to each feature group. It takes the union of outliers sets in all feature groups to declare all outlying observations. Therefore, the number of declared outliers by WATCH is at most $M \times g$.

Accordingly, the WATCH method takes into consideration correlation among categorical attributes by partitioning dimensions into g feature groups. The method aims at detecting hidden outliers (those observations that significantly differ from other observations in subsets of correlated dimensions).

The time complexity of HOT is $\approx O(nq^2 + q|\mathrm{H}|\ max(|he|))$, where $|\mathrm{H}|$ is the total number of hyperedges (frequent itemsets) and $max(|he|)$ is the maximum number of observations found in a single hyperedge. The HOT has a linear complexity with respect to the number of observations and the size of the hyper graph, while, it has a quadratic complexity with respect to the number of categorical variables.

The time complexity of k-LOF can be divided into two parts: Building similarity graph G and computing similarity and outlying scores. The time complexity for building the similarity graph is $O(n^2q)$. However, the time complexity of computing similarity and outlying scores is $\approx O(n^2 \sum_{i=1}^{k} i)$. Therefore, the total time complexity of k-LOF is $\approx O(n^2(q + \sum_{i=1}^{k} i))$.

The time complexity of the WATCH algorithm consists of three parts. First, the building feature groups phase requires $O(qgt)$, *where* t is number of iterations. Second, the complexity of detecting outliers in all feature groups requires $O(nsg)$. Third, the complexity of finding the optimal number of feature groups is $O((q/2-1)qgt)$. Therefore, the overall complexity of WATCH is $O(q^2gt/2 + ngs)$.

Accordingly, the density-based anomaly detection methods, HOT, k-LOF, and WATCH are computationally expensive due to the quadratic cost of measuring similarity with respect to the number of observations. Additionally, these methods require specifying decision parameters k and θ in advance. Specifying suitable values of these parameters is difficult and the results are also very sensitive to these parameters.

6 CLUSTERING-BASED METHODS

Clustering-based anomaly detection methods rely on clustering categorical datasets then identify observations that are located in sparse regions as outliers. There are two methods that belong to this category: (a) Ranking-based Outliers Analysis and Detection (ROAD) [167, 169], and (b) the Rough-ROAD method [168, 170].

The Ranking-based Outliers Analysis and Detection (ROAD) method defines two types of outliers; frequency-based and clustering-based outliers. Frequency-based outliers are those observations that have infrequent categories (small average marginal frequencies). However, clustering-based outliers are those observations that have infrequent combinations of frequent categories.

ROAD develops two different ranking schemes; one for each outlier-type. First, it computes a density score for each observation (the average marginal frequencies) as $den(x_i) = \frac{1}{q}\sum_{j=1}^{q} f(x_{ij})$, which is the same as $\mathrm{AVF}(x_i)$ in Equation (1). ROAD sorts observations based on their density

scores. Then it gives the higher probability for being frequency-based outliers to those observations having small density scores.

Second, ROAD partitions the given categorical dataset into k clusters using the $k-$mode algorithm [92]. Then it defines the set of big clusters, BC, as the clusters that contain at least $\alpha\%$ of observations, where α is the big cluster threshold. Additionally, ROAD computes the distance between an observation x_i and a certain cluster C_l, $d(x_i, C_l)$, as $d(x_i, C_l) = \sum_{j=1}^{q} \phi(x_{ij}, z_{lj})$, where z_{lj} is the jth category in the representative (mode), z_l, in the lth cluster, C_l, and $\phi(x_{ij}, z_{lj})$ is computed as

$$\phi(x_{ij}, z_{lj}) = \begin{cases} 1, & \text{if } x_{ij} \neq z_{lj}, \\ 1 - \frac{|C_{lj}|}{|C_l|}, & \text{otherwise}, \end{cases} \tag{38}$$

where $|C_{lj}|$ is the number of observations that have x_{ij} and $|C_l|$ is the number of observations in the lth cluster.

ROAD defines a cluster-based ranking scheme, $clus\text{-}rank(x_i)$ based on the distance to the nearest big cluster. The larger distance to nearest big cluster, the higher probability for being cluster-based outliers. ROAD defines two sets of most likely outliers by labeling top M observations in each rank. For the sake of integration, ROAD defines an observation as an outlier if it belongs to frequency-based outliers or cluster-based outliers. It computes the union of the sets of outliers. Accordingly, the ROAD takes into account marginal frequencies as well as joint frequencies. However, it requires two parameters; M and the big cluster threshold, α.

The Rough-ROAD method extends the ROAD method. Here, the k-modes algorithm [92] is modified to the rough k-modes algorithm based on using rough sets theory [136] to capture the uncertainty and deal with ambiguity regarding the membership of outliers to certain clusters. Rough-ROAD consists of two phases; rough clustering phase and ranking outliers phase. The only difference between ROAD and Rough-ROAD is in the clustering phase.

Similar to the original k-modes algorithm, the rough k-modes aims at partitioning categorical data into k clusters. Each cluster, C_j, has a representative called C_j's mode, Z_j, the observation with the highest density in C_j. The rough k-modes consists of three basic steps: the modes random initialization step, the assigning of each observation to clusters, and the updating of the modes based on the assigning step. Rough k-modes and k-modes are iterative algorithms repeating the second and the third steps until convergence (no change in clusters' distributions).

For each cluster, C_j, rough k-modes defines two sets of observations a lower approximation, $\underline{C_j}$, and an upper approximation, $\overline{C_j}$. The lower approximation $\underline{C_j}$ contains all objects that are surely located in cluster C_j. In other words, it contains all objects, where C_j is the only close cluster to these objects. Whereas, the upper approximation $\overline{C_j}$ contains all objects that are close to two or more clusters. After the mode initialization step, the Rough-ROAD method assigns an object to either a single lower approximation set of a rough cluster or multiple upper approximation sets of rough clusters. For each object, x_i, it identifies the nearest cluster, C_j, which has the minimum $d(x_i, Z_j)$, where $1 \leq j \leq k$. Then, it looks for the set of clusters, T, that are close with respect to its nearest cluster, that is,

$$T = \left\{ C_l : \frac{d(x_i, Z_l)}{d(x_i, Z_j)} \leq \varepsilon \right\}, \tag{39}$$

where $\varepsilon \geq 1$ is the roughness parameter. If the set of close clusters T is empty, then x_i has only one close rough cluster C_j. Thus, x_i is assigned only to the lower approximation of the jth cluster, $\underline{C_j}$. However, if T is not empty, then x_i has two or more close rough clusters C_j and C_l. Thus, x_i is assigned to the upper approximation of those clusters $\overline{C_j}$ and $\overline{C_l}$.

After Rough-ROAD assigns all objects to rough clusters, it recalculates the new clusters' modes, $Z_1, Z_2, \ldots, Z_k$. In each rough cluster, C_j, it searches for the object that has the maximum density to be the mode, Z_j. Then it iterates assigning objects to rough clusters and updating clusters' modes until convergence (there is no change in the clusters).

After the rough clusters are built, outliers detection in Rough-ROAD is the same as the ROAD method. It is based on building a ranking scheme for each type of outliers.

Accordingly, the Rough-ROAD is based on the rough set theory to capture the uncertainty and deal with ambiguity regarding the dataset taking into account the density of objects within its cluster as well as the marginal frequencies. However, it requires the following five parameters: the number of outliers M, the roughness parameter ε, the lower approximation weight parameter w_l, the number of clusters k, and the big cluster threshold α.

The time complexity of ROAD consists of three parts. First, building the frequency-based ranking scheme requires $O(nqc_{max})$, where c_{max} is maximum number of categories per categorical variable. Second, the complexity of computing the cluster-based ranking scheme needs $O(nqk^2 + nqkt)$, where k is the number of clusters and t is the number of iterations required for convergence in the $k-$mode algorithm. Third, the complexity of ranking observation is $O(n\ log\ n)$. Therefore, the overall complexity of ROAD is $O(nqc_{max} + nqk^2 + nqkt + n\ log\ n)$. The time complexity of Rough-ROAD is as complex as that of ROAD. Therefore, the clustering-based methods are computationally expensive.

7 DISTANCE-BASED METHODS

The distance-based anomaly detection approach for categorical data extends the concept of distance-based anomalies for quantitative data [16, 103, 104, 142]. There are many definitions of anomalies in distance-based approach. These include:

- Anomalies are the M observations whose average distances to the k nearest neighbors are the greatest [16, 17].
- Anomalies are the M observations whose distances to the kth nearest neighbor are the greatest [142].
- Anomalies are the observations that have fewer than p observations within a certain distance d [103, 104].
- Anomalies are observations that have the highest z-scores of the average distances to the k-nearest neighbors. That is, first we compute the average distance of each observation to its k-nearest neighbors. Then, we standardized these average distances and obtain the z-scores of the average distances. The observations with z-scores greater than a threshold θ (e.g., 3) are declared as outliers [63]. This method does not require the parameter M in advance, but it assumes that the z-scores follow a standard normal distribution to help in choosing value of θ.

Examples of distance-based methods for categorical data are (a) Orca (name of software) [24], (b) a method called iOrca [28], (c) the Common Neighbor-based distance (CNB) [113], and (d) the Recursive Binning and Re-Projection (RBRP) method [74].

Orca identifies anomalies in mixed datasets. The categorical part of Orca uses the Hamming distance to compute the distance between two categorical observations. The Hamming distance between two categorical observations is the number of mismatches between them [93]. The output of Hamming distance is an integer value between 0 and q, the number of categorical variables in the data. For each observation x_i, Orca computes an outlying score that is the average Hamming distance between x_i and its k nearest neighbors. Orca declares M observations with the highest outlying scores as outliers.

However, a method called indexed Orca (iOrca) [28] is proposed to speed up Orca. iOrca depends on a simple index technique by choosing a random observation, x_r, most likely to be inlier, and then rank all observations based on the distance from x_r. Then, observations are assessed based on their distances to x_r. iOrca can be up to an order of magnitude faster than Orca but in the worst case (when the randomly chosen reference observation is an outlier) iOrca becomes worse than Orca.

The CNB is similar to Orca but it relies on similarity-based distances [40] instead of the Hamming distances. The similarity-based distance between a pair of objects x_i and x_m, $d_{\text{CNB}}(x_i, x_m, \theta)$, is computed in steps. First, finding the common neighbors set

$$\text{CNS}(x_i, x_m, \theta) = \text{NS}(x_i) \cap \text{NS}(x_m), \tag{40}$$

where $\text{NS}(x_i)$ is the set of nearest neighbors for (x_i), which contains an object, x_r, that has $sim(x_i, x_r) \geq \theta$, where $sim(x_i, x_r)$ is the number of matched categories and θ is a predetermined parameter. Second, the distance between x_i and x_m, $d_{\text{CNB}}(x_i, x_m, \theta)$, is computed as

$$d_{\text{CNB}}(x_i, x_m, \theta) = 1 - \left(\frac{\log_2 \mid \text{CNS}(x_i, x_m, \theta) \mid}{\log_2 n} \right). \tag{41}$$

CNB declares the M observations whose average distances to the k nearest neighbors are the highest as anomalies.

The time complexity of Orca is $O(n^2 q)$ in the worst case, whereas the time complexity of CNB is $O(n^2(k + S(\theta) + q) + n(k + M))$, where $S(\theta)$ is the average number of neighbors.

The distance-based anomaly detection methods are not efficient for very large datasets due to its quadratic time complexity with respect to the number of observations [24, 28, 74]. Orca performs a pruning strategy to improve the performance of the distance-based anomaly detection methods. For each observation, x_i, if its score during the calculation is less than a certain threshold (the smallest score of anomalies in that iteration), then further processing on that observation is not needed, because it can no longer be an outlier. In the worst case (when observations are arranged in ascending order according to Orca scores), Orca still takes a quadratic time. However, the authors mentioned that Orca runs close to linear time in practice [74].

The Recursive Binning and Re-Projection (RBRP) method is proposed in Reference [74] to reduce the time complexity of distance-based anomaly detection methods to $O(nq \log n + qn^2)$, where n is number of observations and q is number of variables. RBRP groups observations in clusters and then, for each observation x_i, it searches for its nearest neighbors in its cluster first then the next closest cluster and so on. In practice, RBRP outperforms Orca by an order of magnitude [74].

Distance-based anomaly detection methods have the following problems. Their results are very sensitive to the values of the parameters θ and k (number of nearest neighbors). Moreover, they require identifying the number of anomalies M in advance, which is impractical in real applications. They ignore the dependency among categorical variables as well as the number of categories c_j. Finally, they have expensive time complexity $O(n^2 q)$ in the worst case. In addition, the outputs of RBRP and iOrca are random, because the initial observation is chosen at random in the case of iOrca and the RBRP uses the K-mode algorithm, the output of which is random.

8 INFORMATION THEORETIC METHODS

The idea behind information theoretic approach relies on the direct relationship between the existence of anomalies and the amount of noise in the dataset. Consequently, the problem of anomaly detection can be transformed into an optimization problem, where the objective is finding the set of anomalies that maximizes the information gain, or minimizing the uncertainty, of the inlier observations. Most of these methods use an entropy as an information gain or as an uncertainty

measure. These methods can be further classified into two groups based on the type of entropy they use: (a) the Shannon entropy and (b) the Holo entropy.

8.1 Methods Based on the Shannon Entropy

The Shannon entropy of a dataset X is defined as [160]

$$\mathcal{E}(X) = -\sum_{x_i \in X} p(x_i) \log p(x_i), \tag{42}$$

where x_i is an observation in X and $p(x_i)$ is the relative frequency of x_i in X. This can be interpreted as the amount of information contained in the dataset.

The methods based on the Shannon entropy search for a set of observations that minimize the Shannon entropy of the remaining observations. Examples of these methods are (a) the Local Search heuristic Algorithm (LSA) [87] and (b) the Greedy Algorithm (GA) [88].

The LSA finds the M observations, thought to be outliers, that minimize the Shannon entropy of $X_{(I)}$ (the dataset X without the M observations indexed by I). The LSA algorithm builds all C_M^n possible combinations of M, computes the entropy $\mathcal{E}(X_{(I)})$ for each combination, and chooses I that gives the smallest value of $\mathcal{E}(X_{(I)})$. The time complexity of LSA is $O(nqC_M^n)$. LSA algorithm scans the dataset C_M^n times. Therefore, it is infeasible for large n and M.

Since LSA algorithm is not applicable for large data due to its huge time complexity, the GA [88] is proposed to decrease the complexity of LSA. It assumes that the observations are independent from each other. The GA searches for the observation that minimizes the entropy of the remaining $(n-1)$ observations, then it declares that observation as an anomaly and removes it from X. It repeats for the remaining observations until M anomalies have been declared. The time complexity of the GA is $O(nqM)$. Therefore, it significantly decreases the number of dataset scans from C_M^n to M. Although, the GA substantially decreases time complexity of LSA algorithm, its time complexity is still high.

The Shannon entropy-based methods, the LSA and the GA, have the following problems. The time complexity is too high especially when the number of observations and the number of anomalies are large. Moreover, they require identifying the number of anomalies in advance, M, which is impractical in real applications. They are based on the joint frequency, which is always very low especially in large datasets (large number of categorical variables). Therefore, the problem of ties happens due to many subsets giving the same minimum value of the entropy. In addition, they are subject to the masking (failing to declare some outliers) and swamping (declaring some inliers as outliers) problems.

8.2 Methods Based on the Holo Entropy

The total correlation,

$$C_t(X) = \sum_{j=1}^{q} \mathcal{E}(X_j) - \mathcal{E}(X), \tag{43}$$

measures the shared information within a dataset. It is the sum of mutual information of all variables in a dataset X. It can be interpreted as a measure of purity or goodness of the dataset. While the Holo entropy is defined as

$$\mathrm{HE}(X) = \mathcal{E}(X) + C_t(X) = \mathcal{E}(X) + \sum_{j=1}^{q} \mathcal{E}(X_j) - \mathcal{E}(X) = \sum_{j=1}^{q} \mathcal{E}(X_j). \tag{44}$$

Consequently, the total correlation as well as the Shannon entropy are aggregated together into one measure to compute information gain in the dataset.

The Holo entropy-based anomaly detection methods search for the M observations, thought to be outliers, that minimize the Holo entropy of $X_{(I)}$ (the dataset X without M observations indexed by I) [140, 179, 180]. To avoid multiple scans for each observation, x_i (as in the GA), the differential Holo entropy is calculated, $h^x(x_i)$, as the change in the Holo entropy between the dataset with and without x_i. The higher differential Holo entropy an observation has, the higher probability of being an outlier. Then the M observations with the highest differential Holo entropy are labeled as outliers. Examples of the Holo entropy-based algorithms are (a) the Information Theoretic-based (ITB) [180] and (b) the Excess Entropy-based (EEB) [152]. Each of these can use either a single pass (SP) or a step by step (SS). Accordingly, we have four algorithms: ITB-SP, ITB-SS, EEB-SP, and EEB-SS.

In the single pass the dataset is scanned only once then observations with the highest outlier factor (differential entropy) are declared as outliers. The step by step means scanning data sets M times. In each time, the observation with the highest outlier factor (differential Holo entropy) is declared as an outlier and we repeat until M observations are declared as outliers.

The difference between ITB and EEB is in the type of entropy they use. The ITB relies on the Holo Entropy, whereas, the EEB is based on computing excess entropy $\mathcal{E}_E$, also known as the dual total correlation or binding information. The excess entropy quantifies the amount of entropy present in a dataset after subtracting the sum of the entropies of each variable conditioned upon all other variables. It is calculated as

$$\mathcal{E}_E(X) = \mathcal{E}(X) - \sum_{j=1}^{q} \mathcal{E}(X_j|(X \setminus X_i)), \tag{45}$$

where $\mathcal{E}(X \setminus X_i)$ is the dataset X with the variable X_i removed.

The time complexity of ITB-SS and EEB-SS is $O(nqM)$ but the time complexity of ITB-SP and EEB-SP is $O(nq)$. Thus, they significantly decrease the number of dataset scans from C_m^n to either one dataset scan in the case of ITB-SP and EEB-SP or to M dataset scans in the case of ITB-SS and EEB-SS. The Holo entropy-based methods, ITB and EEB require identifying the number of anomalies in advance, M. The ITB-SP and EEB-SP scan dataset only once, which means they are fast and scalable but they may have the masking and swamping problems.

9 COMPRESSION-BASED METHODS

Compression algorithms are usually used in the fields of communication and storage rather than in data mining. Recently, however, few compression-based anomaly detection methods are proposed based on the fact that anomalies do not comply with the model suggested by the other points in the data. Accordingly, observations that could not compress well are considered as outliers. These methods look for the best compression model that suits the non-outlying data points. Objects that deviate (have bad compression measures) are highlighted as anomalies [31]. Examples of compression-based anomaly detection methods for categorical data are (a) KRIMP [164] and (b) Comprex [15].

KRIMP makes use of the Minimum Description Length (MDL) compression concept to decide whether a categorical observation was drawn from the training distribution or it will be considered as an outlier. MDL finds the optimal compression model, $\ddot{m}$, from the set of compression models $\mathcal{M}$ that minimizes the length of both the encoded dataset and the compression model. An observation, x_i, is highlighted as an anomaly if $L(x_i|\ddot{m}) > \theta$, where $L(x_i|\ddot{m})$ is the encoded length of x_i by the optimal compression model, $\ddot{m}$, and θ is a decision parameter.

KRIMP is based on the idea of code tables. A code table consists of two columns. The first column contains the itemsets and the second column contains their codes. Itemsets are sorted in descending order according to their lengths then their frequencies. Higher order (that is, longer

and more frequent) itemsets take shorter code. Each observation is represented by a set of non-overlapping itemsets that completely cover all vales of that observation. An anomaly can be seen as an observation that contains infrequent itemsets and hence, it's encoded code is longer than other observations.

Comprex is similar to KRIMP but it uses multiple code tables instead of one. It makes use of correlation among categorical attributes in building a dictionary (code table) for each strongly correlated set of attributes. It measures the variables correlation by the Information Gain (IG). IG is measured as the average number of bits we can save when compressing X_i and X_j together. If the saving equals zero, then it means that no saving from considering X_i and X_j jointly, hence they are independent. The goal is to find the optimal set of dictionaries that could achieve the minimal lossless compression for the given dataset. Comprex declares observations that have long encoded length as outliers. The idea behind using multiple code tables is to improve compression measures. Moreover, Comprex builds code tables directly from the data by using strongly associated variables instead of using itemset. Therefore, it dose not require building itemsets and hence it avoids its sensitive parameters and expensive cost. The authors claim that Comprex is a parameter-free method, but it actually requires a decision parameter, M, the number of anomalies, which have the highest compression cost. Comprex makes use of Cantelli's Inequality [77] to estimate an approximate value of M.

The time complexity of KRIMP is divided into two parts: Finding frequent itemsets and encoding the full dataset. The complexity of finding frequent itemsets by *a priori* algorithm [6] is $\approx O(n(2^Q))$ and the time complexity of encoding the full dataset is $f^2(nQ+1)$, where n is the number of observations, f is the number of frequent itemsets and Q is the total number of categories. Thus, the total time complexity for KRIMP is $\approx O(n(2^Q + f^2(nQ+1)))$. The time complexity of Comprex is $O(nq^2)$, which has a quadratic complexity with respect to the number of categorical variables.

Although compression-based anomaly detection methods for categorical data outperform other relevant methods in terms of error and detection rates, they are computationally very expensive especially for datasets that contain large number of variables. Similar to other anomaly detection methods for categorical data, compression-based algorithms require decision parameters to decide whether an observation is or is not an outlier.

10 SEMISUPERVISED METHODS

Anomaly identification methods can be classified on the basis of the availability of training samples into three approaches: Supervised, Unsupervised, and Semisupervised [91]. Supervised anomaly detection methods may be considered as a binary classification problem, where each observation is categorized into one of two known classes: normal (inliers) and abnormal (anomalies). Supervised anomaly detection methods consist of two phases: The training phase and the testing phase. In the training phase, well-known samples from both classes, normal and abnormal, are required in building a classification model to distinguish between the behavior of anomalous observations and normal ones. The established classification model is then used for predicting the class label for the observations in the testing phase [137].

The unsupervised anomaly detection methods are used when prior information about the class label are not available. These methods try to divide the data into two unbalanced groups, normal and abnormal, based on the belief that normal observations are located in dense regions while abnormal observations are located in sparse regions [35].

The semisupervised anomaly detection methods use samples of normal observations to build a model of normal data in the training phase. Hence, they do not require abnormal samples in the training phase. Then, they classify instances that significantly deviate from the normal model as anomalies [91].

Supervised techniques are usually faster and more accurate than unsupervised techniques, but they require prior information about normal and abnormal observations, which enhances their performances. However, in some applications, the prior information especially for abnormal data are not available or expensive to obtain (e.g., nuclear fault detection). Moreover, different types of anomalies, which did not appear in the training phase, may appear in the testing phase. As a result, semisupervised anomaly detection approach is useful [90].

A commonly used semisupervised anomaly detection methods for quantitative data is the One-class Support Vector Machine (OSVM) [157]. Examples of semisupervised anomaly detection for categorical data are Feature Regression and Classification (FraC) [128, 129] and a Semisupervised Anomaly Detection framework for Categorical data (SAnDCat) [95].

FRaC is a semisupervised feature modeling approach for detecting anomalies in quantitative and categorical data. During the training phase, FRaC makes use of supervised learning algorithms to build ensemble predictive models. In the testing phase, it labels observations that disagree with the predictive models as anomalies. Initially, the set of features (attributes) should be identified for which predictive models will be learnt. A predictive model uses a predefined set of features to predict the values of a certain feature. Next, a supervised learning algorithm is chosen to train feature models such as regression models for quantitative attributes and decision trees for categorical attributes. FRaC may use multiple supervised learning models to predict one feature then it combines them into a feature ensemble model to improve learning accuracy.

In the testing phase, an anomaly score for each observation is computed based on all feature predictors (e.g., the average number of correct predictions per observation). A higher score indicates stronger agreement with the predictive models and a lower score indicates higher probability of being an anomaly. The preceding anomaly score does not take into account uncertainty of classifier feature models. Therefore, FRaC computes the surprisal anomaly score for each observation x_i as

$$\text{FRaC}(x_i|C) = \sum_{j=1}^{D} \text{surprisal}(P(x_{ij}|C_j, \rho_i(x_i))), \tag{46}$$

where C is the set of predictive models, D is number of features for which predictive models are learnt, and $(C_j, \rho_j(x_i))$ is the predicted value of the jth attribute in x_j based on the mapping function $\rho_i(x_i)$ of the predictor C_j. The surprisal anomaly score is interpreted as the amount of evidence that an instance is anomalous [128].

SAnDCat is also a semisupervised anomaly detection framework for categorical data based on Distance Learning for Categorical Attributes (DILCA) [94], which calculates the normalized Euclidean distance between the conditional probability of categories (y_i, y_j), given the values of other attributes called the context attributes X_k. DILCA computes a context-based distance between two categorical values, y_i and y_j, based on the similarity between their probability distribution using the contingency table [94].

SAnDCat consists of two steps. At the beginning, it learns from the training data to build a normality model, $\mathcal{M}$, consisting of one matrix for each categorical attribute, X_j. The elements of this matrix represent the DILCA distance between each pair of categories in the corresponding attribute. Then, SAnDCat computes anomaly score for each instance in the testing data, x_i, as the average distance between x_i and selected k representative instances from the training set. SAnDCat framework includes four heuristics to select k representatives: random, central k instances in the data, closest k instances for each observation, and farthest k instances for each observation. Finally, the SAnDCat either labels M observations having the highest anomaly score as outliers or labels observations with scores exceeding a certain threshold, θ, as outliers [94].

The complexity and accuracy of FRaC is sensitive to the selected predictors that are used to calculate feature models. Some predictors are linear with respect to sample size (e.g., Naive Bayes). Moreover, FRaC executes multiple cross-validation loops for each feature and for each classifier of the ensemble classifiers for each feature. Therefore, the complexity of FRaC in the worst case scenario is $O(n^2q)$ [95]. The time complexity of SAnDCat can be divided into three parts: Building the classification model, choosing k representatives, and making decision. The complexity of building classification model in the training phase using DILCA distance metric is $O(nq^2 \log q)$ [94]. The complexity of choosing the k representatives from the training data in the worst case is $O(n^2)$. The complexity of making decision in the worst case is $O(n \log n)$. Consequently, the time complexity of SAnDCat is $O(nq^2 \log q + n^2 + n \log n)$.

Semisupervised anomaly detection methods for categorical data outperform other methods in terms of detection rate and computational time. Unlike other anomaly detection approaches for categorical data, the semisupervised approaches require normal instances to learn the normal behavior in the training phase.

11 DISCUSSIONS

Existing anomaly detection methods for categorical data face many problems. Three important problems are discussed below:

11.1 Computational Complexity

Time computational complexity is an important issue in data mining research, since most of real applications have huge datasets in terms of the number of observations and the number of categorical variables. Accordingly, the time complexity is an important criterion for choosing anomaly detection methods for categorical data. The time complexities for various methods are given in Table 1.

The methods FPOF, LOADED, ODMAD, FNDI-OD, AA, CA, and KRIMP are based on finding frequent itemsets. They require an exponential time complexity with respect to the number of categorical variables and several passes over the input dataset. Consequently, these techniques are not recommended for datasets with large number of categorical variables.

Distance-based and density-based methods (Orca, CNB, RBRP, iOrca, k-LOF, Indicator Variable, MCA, FRaC, SAnDCat, and MCA using Canberra distance) always take a very long time for distance computations due to nested loops. Measuring pair-wise distances requires a quadratic time complexity with respect to number of observations. Therefore, these techniques are not recommended for datasets with large number of observations.

The time complexities of the methods based on the Shannon entropy, LSA and GA, are $O(nqC_M^n)$ and $O(nqM)$, respectively, and hence LSA is infeasible for large datasets. The GA, ITB-SS, and EEB-SS scan the dataset M times. Consequently, they take very long time especially when the number of observations and number of anomalies are large.

The time complexities of Comprex and CBRW-OD are approximately $O(nq^2)$ and $O(nQ^2 + |E|I_{max} + nq)$, respectively. Therefore, these techniques are not recommended for datasets with large number of categorical variables. The time complexity of ROAD and Rough-ROAD is based on k^2, where k is the number of clusters. The time complexities of ROAD, Rough-ROAD, and WATCH are affected by the number of iterations until convergence (finding good or optimal solution).

The time complexity of AVF, SCF, ITB-SP, EEB-SP, WAVF, WDOD, and CMBOD is $O(nq)$. It linearly increases with respect to the number of observations and the number of categorical variables. Thus, they are efficient in handling huge datasets, but they ignore dependency among categorical variables.

11.2 Human Intervention

Human intervention (input parameters) is another problem for anomaly detection methods for categorical data. Existing methods require one or more input parameters. In real applications, defining the suitable values of these parameters is a hard and critical task. In addition, the results are sensitive to these input parameters. The most commonly required parameters are given in Table 2.

(1) Decision parameters: Existing anomaly detection methods for categorical data compute a score for each observation. To identify whether an observation is an outlier, they require one of the following parameters:
   - M: The number of assumed anomalies in the dataset. Most of existing methods (AVF, SCF, CMBOD, FPOF, FNDI-OD, CBRW-OD, WATCH, ROAD, Rough-ROAD, CNB, Orca, CNB, LSA, GA, ITB-SP, EEB-SP, ITB-SS, EEB-SS, and CompreX) require M in advance.
   - θ: Outlying score threshold. Observations with scores greater than θ are declared as outliers. The methods that use θ are the Indicator Variables method, MCA, WDOD, CA, RHH, AA, HOT, k-LOF, KRIMP, FRaC, and SAnDCaT.
   - *win* and $\triangle S$: If the difference between an observation score and the average of observations scores in its *win* is greater than $\triangle S$, then that observation is declared as an outlier. The parameters *win* and $\triangle S$ are used in LOADED and ODMAD.

(2) Itemset-based parameters: (*minfreq* and *maxlen*) are used in defining frequent itemsets. They are required by FPOF, LOADED, ODMAD, FNDI-OD, HOT, AA, CA, and KRIMP.

(3) k: The number of nearest neighbors in distance-based anomaly detection methods for categorical data. It is required in Bouguessa's method, Orca, CNB, and k-LOF.

11.3 Real and Synthetic Categorical Datasets

Testing the performance of anomaly detection methods requires both real benchmark data as well as synthetic data. Real benchmark quantitative data are abundant. By contrast, there are only few benchmark categorical datasets available such as Breast Cancer and Lymphography datasets [68].

In addition, generating quantitative data with and without anomalies is straightforward, because the definition of anomalies in quantitative data is simple, since anomalies are defined to be a minority of observations that are inconsistent with the pattern suggested by the majority of observations in a dataset. Accordingly, quantitative data with anomalies are generated as follows: The inlier observations are generated from a statistical distribution with certain mean and standard deviation, whereas anomalies are generated from another statistical distribution with different mean and/or standard deviation. The inliers and outliers are then merged together and used for testing the performance of methods for the detection of anomalies in quantitative data. Computation times and detection performance measures (detection rate, precision, and recall) can be computed and used to compare various methods using the same datasets.

By contrast, generating synthetic categorical data and planting anomalies inside them is more difficult than it is for the case for quantitative dataset [36]. The main reason is the lack of agreement in the literature about the definition of outliers. Some anomaly detection papers use synthetic categorical datasets but only to test time performance not detection performance. One way to overcome this difficulty is to first generate quantitative data with outliers, then discretize (categorize) the variables to obtain synthetic categorical data, which can be used to compute time performance and detection performance measures [171].

12 CONCLUSIONS AND FUTURE WORK

Anomaly detection is very important and useful in numerous real-life applications such as identifying computer network intrusions and fraud detection. We surveyed available methods for the identification of anomalies for categorical data in the statistics as well as the machine learning and computer science literatures. There is no general agreement on a single definition of an anomaly in categorical data. We reviewed 36 methods for anomaly detection in categorical data in both the statistics and computer science literatures and classified them into 12 different categories based on the conceptual definition of anomalies they use. We identified the strength and weakness of each method. The computational complexity as well as the required parameters for each method are provided, since they are critical problems in real applications. In addition, we have discussed the common challenges in the detection of anomalies in categorical data.

There are several directions for further research in anomaly detection in categorical data. Extending anomaly detection for categorical data to new research areas such as recommendation systems [56], categorical data streams [39, 124], moving objects [71, 116], and information network [5, 13, 69] may be studied in future research. Defining an automatic critical values instead of specifying the number of anomalies in advance is another direction for future research.

REFERENCES

[1] Abror Abduvaliyev, Al-Sakib Khan Pathan, Jianying Zhou, Rodrigo Roman, and Wai-Choong Wong. 2013. On the vital areas of intrusion detection systems in wireless sensor networks. *IEEE Commun. Surveys Tutor.* 15, 3 (2013), 1223–1237.

[2] Hala Abukhalaf, Jianxin Wang, and Shigeng Zhang. 2015. Outlier detection techniques for localization in wireless sensor networks: A survey. *Int. J. Future Gen. Commun. Netw.* 8, 6 (2015), 99–114.

[3] Charu C. Aggarwal. 2017. *Outlier Analysis*, 2nd ed. Springer, Cham.

[4] Charu C. Aggarwal and Philip S. Yu. 2001. Outlier detection for high dimensional data. In *Proceedings of the ACM International Conference on Management of Data (SIGMOD'01)*. 37–46.

[5] Charu C. Aggarwal, Yuchen Zhao, and Philip S. Yu. 2011. Outlier detection in graph streams. In *Proceedings of the ACM IEEE International Conference on Data Engineering (ICDE'11)*. 399–409.

[6] Rakesh Agrawal and Ramakrishnan Srikant. 1994. Fast algorithms for mining association rules in large databases. In *Proceedings of International Conference on Very Large Data Bases (VLDB'94)*. 487–499.

[7] A. Agresti. 2010. *Analysis of Ordinal Categorical Data* (2nd ed.). John Wiley & Sons, New York, NY.

[8] A. Agresti. 2013. *Categorical Data Analysis* (3rd ed.). John Wiley & Sons, New York, NY.

[9] Malik Agyemang, Ken Barker, and Rada Alhajj. 2006. A comprehensive survey of numeric and symbolic outlier mining techniques. *Intell. Data Anal.* 10(6) (2006), 521–538.

[10] Mohiuddin Ahmed, Abdun Naser Mahmood, and Jiankun Hu. 2016. A survey of network anomaly detection techniques. *Netw. Comput. Appl.* 60 (2016), 19–31.

[11] Mohiuddin Ahmed, Abdun Naser Mahmood, and Md. Rafiqul Islam. 2016. A survey of anomaly detection techniques in financial domain. *Future Gen. Comput. Syst.* 55 (2016), 278–288.

[12] P. Ajitha and E. Chandra. 2015. A survey on outliers detection in distributed data mining for big data. *J. Basic Appl. Sci. Res.* 5, 2 (2015), 31–38.

[13] Leman Akoglu, Mary Mcglohon, and Christos Faloutsos. 2010. OddBall: Spotting anomalies in weighted graphs. In *Proceedings of the Pacific Asia Knowledge Discovery and Data Mining (PAKDD'10)*. 420–431.

[14] Leman Akoglu, Hanghang Tong, and Danai Koutra. 2015. Graph-based anomaly detection and description: A survey. *Data Min. Knowl. Discov.* 29, 3 (2015), 626–688.

[15] Leman Akoglu, Hanghang Tong, Jilles Vreeken, and Christos Faloutsos. 2012. Fast and reliable anomaly detection in categorical data. In *Proceedings of the ACM International Conference on Information and Knowledge Management, (CIKM'12)*. 415–424.

[16] Fabrizio Angiulli, Stefano Basta, and Clara Pizzuti. 2006. Distance-based detection and prediction of outliers. *IEEE Trans. Knowl. Data Eng.* 18(2) (2006), 145–160.

[17] Fabrizio Angiulli and Fabio Fassetti. 2002. Fast outlier detection in high dimensional spaces. In *Proceedings of the European Conference on the Principles of Data Mining and Knowledge Discovery*. 19–26.

[18] Yagnik N. Ankur and Ajay Shanker Singh. 2014. Oulier analysis using frequent pattern mining: A review. *Int. J. Comput. Sci. Info. Technol.* 5, 1 (2014), 47–50.

[19] N. Archana and S. S. Pawar. 2014. Survey on outlier pattern detection techniques for time-series data. *Int. J. Sc. Res.* 1, 1 (2014), 1852–1856.

[20] Tony Bailetti, Mahmoud Gad, and Ahmed Shah. 2016. Intrusion learning: An overview of an emergent discipline. *Technol. Innovat. Manage. Rev.* 6, 2 (2016), 15–20.

[21] U. A. B. U. A. Bakar, Hemant Ghayvat, S. F. Hasanm, and S. C. Mukhopadhyay. 2016. Activity and anomaly detection in smart home: A survey. In *Next Generation Sensors and Systems*, Subhas Chandra Mukhopadhyay (Ed.). Springer, New York, NY, Chapter 9, 191–220.

[22] Zuriana Abu Bakar, Rosmayati Mohemad, Akbar Ahmad, and Mustafa Mat Deris. 2006. A comparative study for outlier detection techniques in data mining. In *Proceedings of IEEE International Conference on Cybernetics and Intelligent Systems.* 1–6.

[23] V. Barnett and T. Lewis. 1994. *Outliers in Statistical Data* (3rd ed.). John Wiley & Sons, New York, NY.

[24] S. Bay and M. Schwabacher. 2003. Mining distance-based outliers in near linear time with randomization and a simple pruning rule. In *Proceedings of the ACM International Conference on Knowledge Discovery and Data Mining, SIGKDD.* 29–38.

[25] Eric J. Beh. 2008. Simple correspondence analysis of nominal-ordinal contingency tables.*J. Appl. Math. Decis. Sci.* 228 (2008), 1–17.

[26] Alka P. Beldar and Vinod S. Wadne. 2015. The detail survey of anomaly/outlier detection methods in data mining. *Int. J. Multidisc. Curr. Res.* 3 (2015), 462–472.

[27] Clauber Gomes Bezerra, Bruno Sielly Jales Costa, Luiz Affonso Guedes, and Plamen Parvanov Angelov. 2015. A comparative study of autonomous learning outlier detection methods applied to fault detection. In *Proceedings of the IEEE International Conference on Fuzzy Systems (FUZZ-IEEE'15).* 1–7.

[28] Kanishka Bhaduri, Bryan L. Matthews, and Chris R. Giannella. 2011. Algorithms for speeding up distance-based outlier detection. In *Proceedings of the ACM International Conference on Knowledge Discovery and Data Mining, (SIGKDD'11).* 895–867.

[29] Umale Bhagyashree and M. Nilav. 2014. Overview of k-means and expectation maximization algorithm for document clustering. In *Proceedings of the International Conference on Quality Up-gradation in Engineering, Science and Technology (ICQUEST'14).* 5–8.

[30] N. Billor, Ali S. Hadi, and P. Velleman. 2000. Blocked adaptive computationally-efficient outlier nominators. *Comput. Stat. Data Anal.* 34 (2000), 279–298.

[31] Christian Böhm, Katrin Haegler, Nikola S Müller, and Claudia Plant. 2009. CoCo: Coding cost for parameter-free outlier detection. In *Proceedings of the ACM International Conference on Knowledge Discovery and Data Mining, (SIGKDD'09).* 149–158.

[32] Shyam Boriah, Varun Chandola, and Vipin Kumar. 2008. Similarity measures for categorical data: A comparative evaluation. In *Proceedings of the International SIAM Data Mining Conference (SDM'08).* 243–254.

[33] Mohamed Bouguessa. 2014. A mixture model-based combination approach for outlier detection. *Int. J. Artific. Intell. Tools* 23, 4 (2014), 1–21. DOI: https://doi.org/10.1142/S0218213014600215

[34] Mohamed Bouguessa. 2015. A practical outlier detection approach for mixed-attribute data. *Expert Syst. Appl.* 42 (2015), 8637–8649.

[35] M. M. Breunig, H. Kriegel, R. T. Ng, and J. Sander. 2000. LOF: Identifying density–based local outliers. In *Proceedings of the ACM International Conference on Management of Data (SIGMOD'00).* 93–104.

[36] 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 Min. Knowl. Discov.* 30, 4 (2016), 891–927.

[37] E. Castillo, J. M. Gutiérrez, and A. S. Hadi. 1997. *Expert Systems and Probabilistic Network Models.* Springer-Verlag, New York, NY.

[38] V. Chandola, Arindam Banerjee, and Vipin Kumar. 2009. Anomaly detection: A survey. *ACM Comput. Surveys* 41(3) (2009), 1–58. DOI: https://doi.org/10.1145/1541880.1541882

[39] V. Chandola, A. Banerjee, and V. Kumar. 2012. Anomaly detection for discrete sequences: A survey. *Trans. Knowl. Data Eng.* 24(5) (2012), 823–839.

[40] V. Chandola, S. Boriah, and V. Kumar. 2008. *Understanding Categorical Similarity Measures for Outlier Detection.* Technical Report. University of Minnesota, Department of Computer Science and Engineering, 1-46.

[41] V. Chandola, S. Boriah, and V. Kumar. 2009. A framework for exploring categorical data. In *Proceedings of the International SIAM Data Mining Conference (SDM'09).* 187–198.

[42] S. Chatterjee and Ali S. Hadi. 1986. Influential observations, high leverage points, and outliers in regression. *Stat. Sci.* 1 (1986), 379–416.

[43] S. Chatterjee and Ali S. Hadi. 1988. *Sensitivity Analysis in Linear Regression.* John Wiley & Sons, New York, NY.

[44] Sanjay Chawla and Pei Sun. 2006. SLOM: A new measure for local spatial outliers. *Knowl. Info. Syst.* 9 (2006), 412–429.

[45] Haibin Cheng, Pang-Ning Tan, Christopher Potter, and Steven A. Klooster. 2009. Detection and characterization of anomalies in multivariate time series. In *Proceedings of the SIAM International Conference on Data Mining (SDM'09)*. 413–424.
[46] HyungJun Cho and Soo-Heang Eo. 2016. Outlier detection for mass spectrometric data. In *Statistical Analysis in Proteomics*, Klaus Jung (Ed.). Springer, New York, NY, Chapter 5, 91–102.
[47] Gregory F. Cooper. 1990. The computational complexity of probabilistic inference using Bayesian belief networks. *Artific. Intell.* 42 (1990), 393–405.
[48] Denis Cousineau and Sylvain Chartier. 2015. Outliers detection and treatment: A review. *Int. J. Psychol. Res.* 3, 1 (2015), 58–67.
[49] J. Vijay Daniel, S. Joshna, and P. Manjula. 2013. A survey of various intrusion detection techniques in wireless sensor networks. *Int. J. Comput. Sci. Mobile Comput.* 2, 9 (2013), 235–246.
[50] K. Das and J. Schneider. 2007. Detecting anomalous records in categorical datasets. In *Proceedings of the ACM International Conference on Knowledge Discovery and Data Mining (SIGKDD'07)*. 220–229.
[51] K. Das, J. Schneider, and D. B. Neill. 2008. Anomaly pattern detection in categorical datasets. In *Proceedings of the ACM International Conference on Knowledge Discovery and Data Mining (SIGKDD'08)*. 169–176.
[52] Dhwani Dave and Tanvi Varma. 2014. A review of various statistical methods for outlier detection. *Int. J. Comput. Sci. Eng. Technol.* 5, 2 (2014), 137–140.
[53] Herv Debar, Marc Dacier, and Andreas Wespi. 1999. Towards a taxonomy of intrusion-detection systems. *Comput. Netw.* 31, 9 (1999), 805–822.
[54] Alfonso Iodice D'Enza and Michael Greenacre. 2012. Multiple correspondence analysis for the quantification and visualization of large categorical data sets. In *Advanced Statistical Methods for the Analysis of Large Data-Sets*, Agostino Di Ciaccio, Mauro Coli, and Jose Miguel Angulo Ibañez (Eds.). Springer, 453–463.
[55] Mr. Mukesh K. Deshmukh and A. S. Kapse. 2016. A survey on outlier detection technique in streaming data using data clustering approach. *Int. J. Engineering and Computer Science* 5, 1 (2016), 15453–15456.
[56] Christian Desrosiers and George Karypis. 2011. A comprehensive survey of neighborhood-based recommendation methods. In *Recommender Systems Handbook*. Springer-Verlag New York, NY, 107–144.
[57] R. Lakshmi Devi and R. Amalraj. 2015. Hubness in unsupervised outlier detection techniques for high dimensional data–A survey. *Int. J. Comput. Appl. Technol. Res.* 4, 11 (2015), 797–801.
[58] Jiten Harishbhai Dhimmar and Raksha Chauhan. 2014. A survey on profile-injection attacks in recommender systems using outlier analysis. *Int. J. Adv. Res. Comput. Sci. Manage. Studies* 2, 12 (2014), 356–359.
[59] Xuemei Ding, Yuhua Li, Ammar Belatreche, and Liam P. Maguire. 2014. An experimental evaluation of novelty detection methods. *Neurocomputing* 135 (2014), 313–327.
[60] K. T. Divya and N. S. Kumaran. 2016. Survey on outlier detection techniques using categorical data. *Int. Res. J. Eng. Technol.* 3 (2016), 899–904.
[61] Paul Dokas, Levent Ertoz, Vipin Kumar, Aleksandar Lazarevic, Jaideep Srivastava, and Pang-Ning Tan. 2002. Data mining for network intrusion detection. In *Proceedings of the NSF Workshop on Next Generation Data Mining*. 21–30.
[62] Jin Du, Qinghua Zheng, Haifei Li, and Wenbin Yuan. 2005. The research of mining association rules between personality and behavior of learner under web-based learning environment. In *Proceedings of the the International Conference on Advances in Web-Based Learning (ICWL'05)*. 15–26.
[63] David Ebdon. 1991. *Statistics in Geography: A Practical Approach-Revised with 17 Programs*. Wiley-Blackwell, Hoboken, NJ.
[64] Syed Masum Emran and Nong Ye. 2001. Robustness of Canberra metric in computer intrusion detection. In *Proceedings of the IEEE Workshop on Information Assurance and Security*. New York, NY, 80–84.
[65] Hadi Fanaee-T and João Gama. 2016. Tensor-based anomaly detection: An interdisciplinary survey. *Knowl-Based Syst.* 98 (2016), 130–147.
[66] Elaine R. Faria, Isabel J. C. R. Goncalves, A. C. P. L. F. de Carvalho, and J. Gama. 2015. Novelty detection in data streams. *Artific. Intell. Rev.* 45, 2 (2015), 235–269.
[67] E. W. Forgy. 1965. Cluster analysis of multivariate data: Efficiency versus interpretability of classifications. *Biometrics* 21 (1965), 768–780.
[68] A. Frank and A. Asuncion. 2018. UCI Machine Learning Repository. Retrieved from http://archive.ics.uci.edu/ml/datasets.html.
[69] Jing Gao, Feng Liang, Wei Fan, Chi Wang, Yizhou Sun, and Jiawei Han. 2010. On community outliers and their efficient detection in information networks. In *Proceedings of the ACM International Conference on Knowledge Discovery and Data Mining (SIGKDD'10)*. 813–822.
[70] Pedro Garcia-Teodoro, J. Diaz-Verdejo, Gabriel Maciá-Fernández, and Enrique Vázquez. 2009. Anomaly-based network intrusion detection: Techniques, systems and challenges. *Comput. Security* 28, 1 (2009), 18–28.

[71] Yong Ge, Hui Xiong, Zhi-Hua Zhou, Hasan Ozdemir, Jannite Yu, and K. C. Lee. 2010. TOP-EYE: Top-k evolving trajectory outlier detection. In *Proceedings of the ACM Conference on Information and Knowledge Management, (CIKM'10)*. 1–4.
[72] Dhiren Ghosh and Andrew Vogt. 2012. Outliers: An evaluation of methodologies. In *Proceedings of the Joint Statistical Meetings*. American Statistical Association, 3455–3460.
[73] A. Ghoting, M. E. Otey, and S. Parthasarathy. 2004. Loaded: Link-based outlier and anomaly detection in evolving data sets. In *Proceedings of the IEEE International Conference on Data Mining (ICDM'04)*. 387–390.
[74] Amol Ghoting, Srinivasan Parthasarathy, and Matthew Eric Otey. 2008. Fast mining of distance-based outliers in high dimensional datasets. *Data Min. Knowl. Discov. J.* 16(3) (2008), 349–364.
[75] Prasanta Gogoi, D. K. Bhattacharyya, Bhogeswar Borah, and Jugal K. Kalita. 2011. A survey of outlier detection methods in network anomaly identification. *Comput. J.* 54, 4 (2011), 570–588.
[76] Gene H. Golub and Charles F. van Loan. 2012. *Matrix Computations*, 3rd ed. John Hopkins University Press.
[77] Geoffrey Grimmett and David Stirzaker. 2001. *Probability and Random Processes*, 3rd ed. Oxford University Press, Oxford, UK.
[78] V. Gunamani and M. Abarna. 2013. A survey on intrusion detection using outlier detection techniques. *Int. J. Sci. Eng. Technol. Res.* 2, 11 (2013), 2063 –2068.
[79] Manish Gupta, Jing Gao, Charu C. Aggarwal, and Jiawei Han. 2014. Outlier detection for temporal data. *Synth. Lect. Data Min. Knowl. Discov.* 5, 1 (2014), 1–129.
[80] Manish Gupta, Jing Gao, Charu C. Aggarwal, and Jiawei Han. 2014. Outlier detection for temporal data: A survey. *IEEE Trans. Knowl. Data Eng.* 26, 9 (2014), 2250–2267.
[81] Ali S. Hadi. 1992. Identifying multiple outliers in multivariate data. *J. Roy. Stat. Soc., Ser. B* 54 (1992), 761–771.
[82] Ali S. Hadi. 1992. A new measure of overall potential influence in linear regression. *Comput. Stat. Data Anal.* 14 (1992), 1–27.
[83] Ali S. Hadi. 1994. A modification of a method for the detection of outliers in multivariate samples. *J. Roy. Stat. Soc., Ser. B* 56 (1994), 393–396.
[84] Ali S. Hadi, A. H. M. Rahmatullah Imon, and Mark Werner. 2009. Detection of outliers. *Wiley Interdisc. Rev.: Comput. Stat.* 1 (2009), 57–70.
[85] Ali S. Hadi and J. S. Simonoff. 1993. Procedure for the identification of outliers in linear models. *J. Amer. Stat. Assoc.* 88 (1993), 1264–1272.
[86] Xiaojuan Han, Yong Yan, Cheng Cheng, Yueyan Chen, and Yanglin Zhu. 2014. Monitoring of oxygen content in the flue gas at a coal-fired power plant using cloud modeling techniques. *IEEE Trans. Instrument. Measure.* 63, 4 (2014), 953–963.
[87] Z. He, X. Xu, and S. Deng. 2005. An optimization model for outlier detection in categorical data. In *Proceedings of the International Conference on Advances in Intelligent Computing*. 400–409.
[88] Z. He, X. Xu, and S. Deng. 2006. A fast greedy algorithm for outlier mining. In *Proceedings of the Pacific Asia Knowledge Discovery and Data Mining (PAKDD'06)*. Singapore, 567–576.
[89] Z. He, X. Xu, J. Z. Huang, and S. Deng. 2005. FP-outlier: Frequent pattern based outlier detection. *Comput. Sci. Info. Syst.* 2 (2005), 726–732.
[90] S. Hido, Y. Tsuboi, H. Kashima, M. Sugiyama, and T. Kanamori. 2011. Statistical outlier detection using direct density ratio estimation. *Knowl. Info. Syst.* 26, 2 (2011), 309–336.
[91] V. J Hodge and J. Austin. 2004. A survey of outlier detection methodologies. *Artific. Intell. Rev.* 22 (2004), 85–126.
[92] Zhexue Huang. 1997. A fast clustering algorithm to cluster very large categorical data sets in data mining. In *Proceedings of the International Data Mining and Knowledge Discovery (DMKM'97), Workshop at the ACM International Conference on Mangagement of Data (SIGKDD)*. 1–8.
[93] Z. Huang and M. K. Ng. 1999. A fuzzy k-modes algorithm for clustering categoircal data. *IEEE Trans. Fuzzy Syst.* 7 (1999), 446–452.
[94] Dino Ienco, Ruggero G. Pensa, and Rosa Meo. 2012. From context to distance: Learning dissimilarity for categorical data clustering. *ACM Trans. Knowl. Discov. Data* 6, 1 (2012), 1–12.
[95] Dino Ienco, Ruggero G. Pensa, and Rosa Meo. 2017. A semisupervised approach to the detection and characterization of outliers in categorical data. *IEEE Trans. Neural Netw. Learn.* 28, 5 (2017), 1017–1029.
[96] Francesca Ieva and Anna Maria Paganoni. 2015. Detecting and visualizing outliers in provider profiling via funnel plots and mixed effect models. *Health Care Manage. Sci.* 18, 2 (2015), 166–172.
[97] ShengYi Jiang, Xiaoyu Song, Hui Wang, Jian-Jun Han, and Qing-Hua Li. 2006. A clustering-based method for unsupervised intrusion detections. *Pattern Recogn. Lett.* 27 (2006), 802–810.
[98] Vineet Joshi and Raj Bhatnagar. 2014. CBOF: Cohesiveness-based outlier factor a novel definition of outlier-ness. In *Proceedings of the International Workshop on Machine Learning and Data Mining in Pattern Recognition (MLDM'14)*. 175–189.

[99] Hossein Joudaki, Arash Rashidian, Behrouz Minaei-Bidgoli, Mahmood Mahmoodi, Bijan Geraili, Mahdi Nasiri, and Mohammad Arab. 2015. Using data mining to detect health care fraud and abuse: A review of literature. *Global J. Health Sci.* 7, 1 (2015), 194–202.

[100] Leonid Kalinichenko, Ivan Shanin, and Ilia Taraban. 2014. Methods for anomaly detection: A survey. In *Proceedings of the All-Russian Conference Digital Libraries: Advanced Methods and Technologies, Digital Collections (RCDL'14).* 20–25.

[101] V. Kathiresan and N. A. Vasanthi. 2015. A survey on outlier detection techniques useful for financial card fraud detection. *Int. J. Innovat. Eng. Technol.* 6, 1 (2015), 226–235.

[102] Ravneet Kaur and Sarbjeet Singh. 2015. A survey of data mining and social network analysis based anomaly detection techniques. *Egypt. Info. J.* 39 (2015), 1–18.

[103] E. M. Knorr, R. T. Ng, and V. Tucakov. 2000. Distance-based outliers: Algorithms and applications. *VLDB J.* 8 (2000), 237–253.

[104] Edwin M. Knorr and Raymond T. Ng. 1997. A unified approach for mining outliers. In *Proceedings of the International Conference of the Centre for Advanced Studies on Collaborative Research (CASCON'97).* 236–248.

[105] A. Koufakou, M. Georgiopoulos, and G. Anagnostopoulos. 2008. Detecting outliers in high-dimensional datasets with mixed attributes. In *Proceedings of the International Conference on Data Mining (DMIN'08).*

[106] A. Koufakou, E. Ortiz, M. Georgiopoulos, G. Anagnostopoulos, and K. Reynolds. 2007. A scalable and efficient outlier detection strategy for categorical data. In *Proceedings of the IEEE International Conference on Tools with Artificial Intelligence (ICTAI'07).* 210–217.

[107] Anna Koufakou, Jimmy Secretan, and Michael Georgiopoulos. 2011. Non-derivable itemsets for fast outlier detection in large high-dimensional categorical data. *Knowl. Info. Syst.* 29, 3 (2011), 697–725.

[108] Aleksandar Lazarevic, Levent Ertöz, Vipin Kumar, Aysel Ozgur, and Jaideep Srivastava. 2003. A comparative study of anomaly detection schemes in network intrusion detection. In *Proceedings of the SIAM International Conference on Data Mining (SDM'03).* 25–36.

[109] Dajiang Lei, Liping Zhang, and Lisheng Zhang. 2013. Cloud model-based outlier detect algorithm for categorical data. *Int. J. Database Theory Appl.* 6, 14 (2013), 199–213.

[110] Deyi Li. 2000. Uncertainty in knowledge representation. *Chinese Eng. Sci.* 2, 10 (2000), 73–79.

[111] Jingchao Li and Jian Guo. 2015. A new feature extraction algorithm based on entropy cloud characteristics of communication signals. *Math. Problems Eng.* 2015 (2015), 1–8.

[112] Junli Li, Jifu Zhang, Ning Pang, and Xiao Qin. 2018. Weighted outlier detection of high-dimensional categorical data using feature grouping. *IEEE Trans. Syst. Man Cybernet.: Syst.* (2018), 1–14.

[113] Shuxin Li, Robert Lee, and Sheau-Dong Lang. 2007. Mining distance-based outliers from categorical data. In *Proceedings of the IEEE International Conference on Data Mining Workshops (ICDM'07).* 225–230.

[114] J. Y. Liang, K. S. Chin, and C. Y. Dang. 2002. A new method for measuring uncertainty and fuzziness in rough set theory. *Int. J. Gen. Syst.* 31 (2002), 331–342.

[115] Song Lin and Donald E. Brown. 2006. An outlier-based data association method for linking criminal incidents. *Decis. Support Syst.* 41 (2006), 604–615.

[116] Wei Liu, Yu Zheng, Sanjay Chawla, Jing Yuan, and Xing Xie. 2011. Discovering spatio-temporal causal interactions in traffic data streams. In *Proceedings of the ACM International Conference on Knowledge Discovery and Data Mining (SIGKDD'11).* 1010–1018.

[117] Xutong Liu, Feng Chen, and Chang-Tien Lu. 2014. On detecting spatial categorical outliers. *GeoInformatica* 18, 3 (2014), 501–536.

[118] Arunanshu Mahapatro and Pabitra Mohan Khilar. 2013. Fault diagnosis in wireless sensor networks: A survey. *IEEE Commun. Surveys Tutor.* 15, 4 (2013), 2000–2026.

[119] Kamal Malik, H. Sadawarti, and G. S. Kalra. 2014. Comparative analysis of outlier detection techniques. *Int. J. Comput. Appl.* 97, 8 (2014), 12–21.

[120] Christopher D. Manning, Prabhakar Raghavan, and Hinrich Schutze. 2008. *Introduction to Information Retrieval.* Cambridge University Press, Cambridge, UK.

[121] José Marinho, Jorge Granjal, and Edmundo Monteiro. 2015. A survey on security attacks and countermeasures with primary user detection in cognitive radio networks. *EURASIP J. Info. Secur.* 2015, 1 (2015), 1–14.

[122] Markos Markou and Sameer Singh. 2003. Novelty detection: A review-part 1: Statistical approaches. *Signal Process.* 83 (2003), 2481–2497.

[123] Markos Markou and Sameer Singh. 2003. Novelty detection: A review-part 2: Neural network based approaches. *Signal Process.* 83 (2003), 2499–2521.

[124] Manoj Mishra and Nitesh Gupta. 2015. To detect outlier for categorical data streaming. *Int. J. Sci. Eng. Res.* 6, 5 (2015), 1–5.

[125] Andrew Moore, Mary Soon Lee, and Brigham Anderson. 1998. Cached sufficient statistics for efficient machine learning with large datasets. *J. Artific. Intell. Res.* 8 (1998), 67–91.
[126] Andrew Moore and W. K. Wong. 2003. Optimal reinsertion: A new search operator for accelerated and more accurate Bayesian network structure learning. In *Proceedings of the 20th International Conference on Machine Learning*. 552–559.
[127] Kazuyo Narita and Hiroyuki Kitagawa. 2008. Detecting outliers in categorical record databases based on attribute associations. In *Progress in WWW Research and Development*. Springer, Berlin, 111–123.
[128] K. Noto, C. Brodley, and D. Slonim. 2010. Anomaly detection using an ensemble of feature models. In *Proceedings of the IEEE International Conference on Data Mining (ICDM'10)*. 953–958.
[129] K. Noto, C. Brodley, and D. Slonim. 2012. FRaC: A feature-modeling approach for semi-supervised and unsupervised anomaly detection. *Data Min. Knowl. Discov.* 25, 1 (2012), 109–133.
[130] Colin O'Reilly, Alexander Gluhak, Muhammad Ali Imran, and Sutharshan Rajasegarar. 2014. Anomaly detection in wireless sensor networks in a non-stationary environment. *IEEE Commun. Surveys Tutor.* 16, 3 (2014), 1413–1432.
[131] M. E. Otey, A. Ghoting, and S. Parthasarathy. 2006. Fast distributed outlier detection in mixed-attribute data sets. *Data Min. Knowl. Discov.* 12, 2–3 (May 2006), 203–228.
[132] Matthew Eric Otey, Srinivasan Parthasarathy, and Amol Ghoting. 2005. An empirical comparison of outlier detection algorithms. In *Proceedings of the International Workshop on Data Mining Methods for Anomaly Detection at ACM International Conference on Knowledge Discovery and Data Mining (SIGKDD'05)*. 1–8.
[133] Guansong Pang, Longbing Cao, and Ling Chen. 2016. Outlier detection in complex categorical data by modeling the feature value couplings. In *Proceedings of the 25th International Joint Conference on Artificial Intelligence*. 1902–1908.
[134] Animesh Patcha and Jung-Min Park. 2007. An overview of anomaly detection techniques: Existing solutions and latest technological trends. *Comput. Netw.* 51(12) (2007), 3448–3470.
[135] M. S. Pawar, D. Amruta, and S. N. Tambe. 2014. A survey on outlier detection techniques for credit card fraud detection. *IOSR J. Comput. Eng.* 16, 2 (2014), 44–48.
[136] Zdzisław Pawlak. 1982. Rough sets. *Int. J. Comput. Info. Sci.* 11, 5 (1982), 341–356.
[137] C. Phua, D. Alahakoon, and V. Lee. 2004. Minority report in fraud detection: Classification of skewed data. *ACM SIGKDD Explor. Newslett.* 6, 1 (2004), 50–59.
[138] Clifton Phua, Vincent C. S. Lee, Kate Smith-Miles, and Ross W. Gayler. 2010. A comprehensive survey of data mining-based fraud detection research. Retrieved from http://arxiv.org/abs/1009.6119.
[139] Marco AF Pimentel, David A Clifton, Lei Clifton, and Lionel Tarassenko. 2014. A review of novelty detection. *Signal Process.* 99 (2014), 215–249.
[140] Srijoni Saha Pradip, Jesica Fernandes Robert, and Jasmine Faujdar Hamza. 2015. Information-theoretic outlier detection for large-scale categorical data. *Int. J. Comput. Sci. Mobile Comput.* 4, 4 (2015), 873–881.
[141] Raghav M. Purankar and Pragati Patil. 2015. A survey paper on an effective analytical approaches for detecting outlier in continuous time variant data stream. *Int. J. Eng. Comput. Sci.* 4, 11 (2015), 14946–14949.
[142] Sridhar Ramaswamy, Rajeev Rastogi, and Kyuseok Shim. 2000. Efficient algorithms for mining outliers from large data sets. In *Proceedings of the ACM International Conference on Management of Data (SIGMOD'00)*. 427–438.
[143] Stephen Ranshous, Shitian Shen, Danai Koutra, Steve Harenberg, Christos Faloutsos, and Nagiza F. Samatova. 2015. Anomaly detection in dynamic networks: A survey. *Wiley Interdisc. Rev.: Comput. Stat.* 7, 3 (2015), 223–247.
[144] Lida Rashidi, Sattar Hashemi, and Ali Hamzeh. 2011. Anomaly detection in categorical datasets using Bayesian networks. In *Proceedings of the 3rd International Conference on Artificial Intelligence and Computational Intelligence, Part II (AICI'11)*. 610–619.
[145] Murad A. Rassam, M. A. Maarof, and Anazida Zainal. 2012. A survey of intrusion detection schemes in wireless sensor networks. *Amer. J. Appl. Sci.* 9, 10 (2012), 1636–1652.
[146] Murad A. Rassam, Anazida Zainal, and Mohd Aizaini Maarof. 2013. Advancements of data anomaly detection research in wireless sensor networks: A survey and open issues. *Sensors* 13, 8 (2013), 10087–10122.
[147] D. Lakshmi Sreenivasa Reddy, B. Raveendra Babu, and A. Govardhan. 2013. Outlier analysis of categorical data using navf. *Informat. Econom.* 17, 1 (2013), 1–5.
[148] Abdolazim Rezaei, Zarinah M. Kasirun, Vala Ali Rohani, and Touraj Khodadadi. 2013. Anomaly detection in online social networks using structure-based technique. In *Proceedings of the International Conference for Internet Technology and Secured Transactions (ICITST'13)*. 619–622.
[149] Ritika, Tarun Kumar, and Amandeep Kaur. 2013. Outlier detection in WSN: A survey. *Int. J. Adv. Res. Comput. Sci. Softw. Eng.* 3, 7 (2013), 609–617.
[150] N. Rokhman, Subanar, and E. Winarko. 2016. Improving the performance of outlier detection methods for Categorical data by using weighting function. *J. Theor. Appl.d Info.n Technol.* 83 (2016), 327–336.
[151] Peter J. Rousseeuw and Katrien Van Driessen. 1998. A fast algorithm for the minimum covariance determinant estimator. *Technometrics* 41 (1998), 212–223.

[152] Ashwini G. Sagade and Ritesh Thakur. 2014. Excess entropy based outlier detection in categorical data set. *Int. J. Adv. Comput. Eng. Netw.* 2, 8 (2014), 56–61.
[153] Aiman Moyaid Said, Dhanapal Durai Dominic, and Brahim Belhaouari Samir. 2013. Outlier detection scoring measurements based on frequent pattern technique. *Res. J. Appl. Sci. Eng. Technol.* 6, 8 (2013), 1340–134.
[154] Arif Sari. 2015. A review of anomaly detection systems in cloud networks and survey of cloud security measures in cloud storage applications. *J. Info. Secur.* 6, 2 (2015), 142–154.
[155] Debajit Sen Sarma and Samar Sen Sarma. 2015. A survey on different graph based anomaly detection techniques. *Indian J. Sci. Technol.* 8, 31 (2015), 1–7.
[156] David Savage, Xiuzhen Zhang, Xinghuo Yu, Pauline Chou, and Qingmai Wang. 2014. Anomaly detection in online social networks. *Soc. Netw.* 39 (2014), 62–70.
[157] Bernhard Schölkopf, John C. Platt, John Shawe-Taylor, Alex J. Smola, and Robert C Williamson. 2001. Estimating the support of a high-dimensional distribution. *Neural Comput.* 13, 7 (2001), 1443–1471.
[158] Junhee Seok and Yeong Seon Kang. 2015. Mutual information between discrete variables with many categories using recursive adaptive partitioning. *Sci. Rep.* 5 (2015), 1–10.
[159] Nauman Shahid, Ijaz Haider Naqvi, and Saad Bin Qaisar. 2015. Characteristics and classification of outlier detection techniques for wireless sensor networks in harsh environments: A survey. *Artific. Intell. Rev.* 43, 2 (2015), 193–228.
[160] Claude Elwood Shannon. 1948. A mathematical theory of communication. *Bell Tele. Syst. Techn. Publ.* 27, 3 (1948), 379–423.
[161] Deep Shikha Shukla, Avinash Chandra Pandey, and Ankur Kulhari. 2014. Outlier detection: A survey on techniques of WSNs involving event and error based outliers. In *Proceedings of the International Conference of Innovative Applications of Computational Intelligence on Power, Energy and Controls with their Impact on Humanity (CIPECH'14).* 113–116.
[162] M. Shyu, K. Sarinnapakorn, I. Kuruppu-Appuhamilage, S. Chen, L. W. Chang, and T. Goldring. 2005. Handling nominal features in anomaly intrusion detection problems. In *Proceedings of the International Workshop on Research Issues in Data Engineering: Stream Data Mining and Applications.* 55–62.
[163] Karanjit Singh and Shuchita Upadhyaya. 2012. Outlier detection: Applications and techniques. *Int. J. Comput. Sci. Iss.* 9, 1 (2012), 307–323.
[164] Koen Smets and Jilles Vreeken. 2011. The odd one out: Identifying and characterising anomalies. In *Proceedings of the SIAM International Conference on Data Mining (SDM'11).* 804–815.
[165] Angela A. Sodemann, Matthew P. Ross, and Brett J. Borghetti. 2012. A review of anomaly detection in automated surveillance. *IEEE Trans. Syst. Man Cybernet., Part C: Appl. Rev.* 42, 6 (2012), 1257–1272.
[166] Garule Supriya and Sharmila M. Shinde. 2015. Outliers detection using subspace method: A survey. *Int. J. Comput. Appl.* 112, 16 (2015), 20–22.
[167] N. N. R. R. Suri, M. N. Murty, and G. Athithan. 2012. An algorithm for mining outliers in categorical data through ranking. In *Proceedings of the 12th IEEE International Conference on Hybrid Intelligent Systems (HIS'12).* 247–252.
[168] N. N. R. R. Suri, M. N. Murty, and G. Athithan. 2013. A rough clustering algorithm for mining outliers in categorical data. In *Proceedings of the 4th International Conference on Pattern Recognition and Machine Intelligence (PReMI'13).* 170–175.
[169] N. N. R. R. Suri, M. N. Murty, and G. Athithan. 2014. A ranking-based algorithm for detection of outliers in categorical data. *Int. J. Hybrid Intell. Syst.* 11 (2014), 1–11.
[170] N. N. R. R. Suri, M. N. Murty, and G. Athithan. 2016. Detecting outliers in categorical data through rough clustering. *Nat. Comput.* 15 (2016), 385–394.
[171] Ayman Taha and Ali S. Hadi. 2013. A general approach for automating outliers identification in categorical data. In *Proceedings of the ACS/IEEE International Conference on Computer Systems and Applications (AICCSA'13).* 1–8.
[172] Ayman Taha and Ali S. Hadi. 2016. Pair-wise association for categorical and mixed attributes. *Info. Sci.* 346 (2016), 73–89.
[173] Ayman Taha and Osman Hegazy. 2010. A proposed outliers identification algorithm for categorical data sets. In *Proceedings of International Conference on Informatics and Systems (INFOS'10).* 1–5.
[174] Yun Wang. 2008. *Statistical Techniques for Network Security: Modern Statistically-Based Intrusion Detection and Protection.* IGI Global, New York, NY.
[175] Yibo Wang and Wei Xu. 2018. Leveraging deep learning with LDA-based text analytics to detect automobile insurance fraud. *Decis. Support Syst.* 105 (2018), 87–95.
[176] Li Wei, Weining Qian, Aoying Zhou, Wen Jin, and Jeffrey X. Yu. 2003. Hypergraph-based outlier test for categorical data. In *Proceedings of the ACM International Conference on Knowledge Discovery and data Mining (SIGKDD'03).* 399–410.
[177] David J. Weller-Fahy, Brett J. Borghetti, and Angela A. Sodemann. 2015. A survey of distance and similarity measures used within network intrusion anomaly detection. *IEEE Commun. Surveys Tutor.* 17, 1 (2015), 70–91.

[178] Jarrod West and Maumita Bhattacharya. 2016. Intelligent financial fraud detection: A comprehensive review. *Comput. Secur.* 57 (2016), 47–66.
[179] Shu Wu and Shengrui Wang. 2011. Parameter-free anomaly detection for categorical data. *Machine Learning and Data Mining in Pattern Recognition. Lecture Notes in Computer Science* 6871 (2011), 112–126.
[180] Shu Wu and Shengrui Wang. 2013. Information-theoretic outlier detection for large-scale categorical data. *IEEE Trans. Knowl. Data Eng.* 25, 3 (2013), 589–602.
[181] Warusia Yassin, Nur Izura Udzir, Zaiton Muda, and Nasir Sulaiman. 2013. Anomaly-based intrusion detection through k-means clustering and naives Bayes classification. In *Proceedings of the International Conference on Computing and Informatics (ICOCI'13).* 298–303.
[182] Jeffrey Xu Yu, Weining Qian, Hongjun Lu, and Aoying Zhou. 2006. Finding centric local outliers in categorical/numerical spaces. *Knowl. Info. Syst.* 9 (2006), 309–338.
[183] Rose Yu, Huida Qiu, Zhen Wen, Ching-Yung Lin, and Yan Liu. 2016. A survey on social media anomaly detection. Retrieevd from http://arxiv.org/pdf/1601.01102.
[184] Ji Zhang. 2013. Advancements of outlier detection: A survey. *ICST Trans. Scal. Info. Syst.* 13, 1 (2013), 1–26.
[185] Yang Zhang, Nirvana Meratnia, and Paul Havinga. 2010. Outlier detection techniques for wireless sensor networks: A survey. *IEEE Commun. Surveys Tutor.* 12, 2 (2010), 159–170.
[186] Xingwang Zhao, Jiye Liang, and Fuyuan Cao. 2014. A simple and effective outlier detection algorithm for categorical data. *Int. J. Mach. Learn. Cybernet.* 5 (2014), 469–477.
[187] Wobbe P. Zijlstra, L. Andries van der Ark, and Klaas Sijtsma. 2011. Outliers in questionnaire data: Can they be detected and should they be removed. *J. Edu. Behav. Stat.* 36 (2011), 186–212.
[188] Arthur Zimek, Erich Schubert, and Hans-Peter Kriegel. 2012. A survey on unsupervised outlier detection in high-dimensional numerical data. *Stat. Anal. Data Min.* 5, 5 (2012), 363–387.

Received January 2017; revised January 2019; accepted January 2019