INTERLIBRARY LOAN

Warning: this work may be protected by the copyright
laws of the United States, Title 17, United States Code.

The WSU Libraries' goal is to provide excellent customer service. Let us know how we are doing by responding to this short survey:
https://libraries.wsu.edu/access_services_survey

Rapid #: -16903425

CROSS REF ID: **1321594**

LENDER: **TAUESE :: Main Library**

BORROWER: **NTE :: Main Library**

TYPE:	Book Chapter
BOOK TITLE:	Outlier Detection: Techniques and Applications
USER BOOK TITLE:	Outlier Detection: Techniques and Applications
CHAPTER TITLE:	Outlier Detection in Categorical Data
BOOK AUTHOR:	Athithan,
EDITION:	
VOLUME:	
PUBLISHER:	
YEAR:	2019
PAGES:	69-93
ISBN:	9783030051273
LCCN:	
OCLC #:	

Processed by RapidX: 11/30/2020 9:32:02 AM

This material may be protected by copyright law (Title 17 U.S. Code)

Chapter 5
Outlier Detection in Categorical Data

Abstract This chapter delves on a specific research issue connected with outlier detection problem, namely *type of data attributes*. More specifically, the case of analyzing data described using categorical attributes/features is presented here. It is known that the performance of a detection algorithm directly depends on the way outliers are perceived. Typically, categorical data are processed by considering the occurrence frequencies of various attributes values. Accordingly, the objective here is to characterize the deviating nature of data objects with respect to individual attributes as well as in the joint distribution of two or more attributes. This can be achieved by defining the measure of deviation in terms of the attribute value frequencies. Also, cluster analysis provides valuable insights on the inherent grouping structure of the data that helps in identifying the deviating objects. Based on this understanding, this chapter presents algorithms developed for detection of outliers in categorical data.

5.1 Introduction

As stated earlier, the aim of this study is to identify data objects that deviate significantly from the rest of the data. Detection of such deviating objects or rare objects and exceptions in data is the purpose of outlier analysis. This has numerous applications in various domains such as fraud detection in financial transactions and intrusion detection in computer networks, etc.

The type of attributes describing data is a key issue among the ones identified connected with outlier detection. More specifically, data described using *qualitative (categorical) attributes* is the focus of this study. The importance of dealing with categorical attributes is evident from the observation that data pertaining to several real world applications are described using such attributes. This emphasizes the need for developing suitable algorithms for detecting outliers in categorical data.

Copyright © 2019. Springer International Publishing AG. All rights reserved.

© Springer Nature Switzerland AG 2019

N. N. R Ranga Suri et al., *Outlier Detection: Techniques and Applications*,
Intelligent Systems Reference Library 155,
https://doi.org/10.1007/978-3-030-05127-3_5

Ranga, Suri, N. N. R., et al. Outlier Detection: Techniques and Applications, Springer International Publishing AG, 2019.
ProQuest Ebook Central, http://ebookcentral.proquest.com/lib/tau/detail.action?docID=5631525.
Created from tau on 2020-11-30 08:29:08.

5.1.1 Categorical Attributes

The scale of measure indicates the nature of information within the values assigned to variables/attributes. There are four scales of measurement popularly used: nominal, ordinal, interval and ratio.

Nominal scale differentiates various values based only on their names. Examples of this type are color, direction, language, etc. Mathematical operations permitted on this type of values and their central tendency are listed in Table 5.1.

Ordinal scale allows to rank order the values to produce a sorted sequence. It includes dichotomous values (such as 'true', 'false') and also non-dichotomous data comprising a spectrum of values (such as 'excellent', 'very good', 'good', 'satisfied', 'poor' for expressing one's opinion).

Interval scale allows for measuring the degree of difference between two values. For example, location information of places indicated using Cartesian coordinates can be used to determine which place is closer to a specified place. Mode, median and arithmetic mean are the central tendency measures applicable on this type of data.

Ratio scale measures values by estimating the ratio between a magnitude of a continuous quantity and a basic unit of the same type. All statistical measures can be computed on this type of data as necessary mathematical operations are defined for this data, as indicated in Table 5.1.

A categorical variable/attribute, belonging to the nominal scale, takes one of a limited number of possible values. The central tendency of the values of a categorical variable is given by its mode. Though numeric values may appear corresponding to a categorical variable, they each represent a logically separate concept and cannot be processed as numbers. In computer science, categorical variables are referred to as enumerated types. The term categorical data applies to data sets that contain one or more categorical variables. Categorical data are also known as nominal or qualitative multi-scale data in different application contexts.

A categorical variable that can take on exactly two values is known as dichotomous (binary) variable. The ones with more than two possible values are called as polytomous variables. Regression analysis on categorical variables is accomplished through multinomial logistic regression.

The acceptable values of a categorical (qualitative) attribute are represented by various categories, as illustrated in Table 5.2. The information on the occurrence

Table 5.1 Details on various scales of measure

Scale	Property	Mathematical operations	Central tendency
Nominal	Membership	$=, \neq$	Mode
Ordinal	Comparison	$>, <$	Median
Interval	Difference	$+, -$	Mean
Ratio	Magnitude	x, /	Geometric mean

Copyright © 2019. Springer International Publishing AG. All rights reserved.

Ranga, Suri, N. N. R., et al. Outlier Detection: Techniques and Applications, Springer International Publishing AG, 2019.
ProQuest Ebook Central, http://ebookcentral.proquest.com/lib/tau/detail.action?docID=5631525.
Created from tau on 2020-11-30 08:29:08.

Table 5.2 Example categorical attributes with sample values

Attribute name	Categories (acceptable values)
Color	Green, red, blue, …
Direction	North, east, west, south
Weather	Sunny, cloudy, raining, …
…	…

frequencies of various categories of a categorical attribute in data is demanded by many data-dependent tasks such as outlier detection.

As already known, the characteristics of a categorical data set can be summarized using the following details.

1. Size of the data set (n)
2. Number of attributes (m)
3. Number of values taken by each attribute ($|DOM(A_r)|, \forall r$)
4. Frequencies of various values of each categorical attribute ($freq(x_{i,r}), \forall i, r$)

It is possible to define any computational method on categorical attributes using the above abstract model.

The simplest way to find similarity between two categorical attributes is to employ the *overlap* measure (Eq. 4.1). This measure treats two categorical attributes as similar if their values are identical, and dissimilar otherwise. A drawback with this measure is that it doesn't distinguish between the different values taken by an attribute. All matches, as well as mis-matches are treated as equal.

Referring to the sample data set shown in Table 5.3, it is obvious that certain combination of the values of the categorical attributes $\{Model, Color\}$ describing the data are more frequent than the others. Bringing in the frequency details into account, it is possible to determine which of the objects are more similar and also the degree of their similarity.

Table 5.3 Frequency distribution of a sample 2-D car data (for illustration)

Model/Color	Black	White	Grey
Maruti-Swift	20	35	45
Tata-Sumo	10	75	15
Honda-City	40	25	35
Toyota-Corolla	35	25	40
Volkswagen-Polo	25	20	55

Copyright © 2019. Springer International Publishing AG. All rights reserved.

Ranga, Suri, N. N. R., et al. Outlier Detection: Techniques and Applications, Springer International Publishing AG, 2019.
ProQuest Ebook Central, http://ebookcentral.proquest.com/lib/tau/detail.action?docID=5631525.
Created from tau on 2020-11-30 08:29:08.

5.1.2 Challenges with Categorical Data

Though there exist a number of methods for outlier detection in numerical data, only a limited number of them can process the data represented using categorical attributes. Outlier detection in categorical data is an evolving problem due to the computational challenges posed by the categorical attributes/features. The fundamental issue in this regard is the difficulty in defining a proximity measure over the categorical values. This is due to the fact that the various values that a categorical variable can assume are not inherently ordered. As a result, many data mining tasks such as determining the nearest neighbor of a categorical object turn out to be non-trivial. Active research happening in this direction is indicative of the importance associated with this aspect. Depending on the application context, supervised measures determine similarity based on class information, while data-driven measures determine the same based on the data distribution.

Many a time, when dealing with data having categorical attributes, it is assumed that the categorical attributes could be easily mapped into numeric values. However, there are instances of categorical attributes, where mapping to numerical attributes is not a straightforward process, and the results greatly depend on the mapping that is used. Consequently, methods such as those based on distance or density measurements are not acceptable, requiring necessary modification in their formulations.

5.2 Clustering Categorical Data

Clustering is an important data mining task required in numerous real life applications. Due to the difficulties in analyzing categorical data, as discussed in the previous section, performing cluster analysis on such data turns out to be an involved activity. However, as an unsupervised learning task, it helps in understanding the inherent grouping structure of the data for the purpose of outlier detection. In that sense, cluster analysis assumes significance towards detecting outliers in the data.

Prior research in this direction has brought out some interesting algorithms for clustering categorical data. A glimpse of some of these algorithms is provided below to facilitate effective data analysis.

5.2.1 ROCK Algorithm

RObust Clustering using linKs (ROCK) algorithm works based on the notion of links between data objects, rather than applying any metric to measure the similarity. The number of links between a pair of objects is given by the number of common neighbors of the objects. The rationale is that data objects belonging to single cluster will in general have a large number of common neighbors, and consequently more

Copyright © 2019. Springer International Publishing AG. All rights reserved.

Ranga, Suri, N. N. R., et al. Outlier Detection: Techniques and Applications, Springer International Publishing AG, 2019.
ProQuest Ebook Central, http://ebookcentral.proquest.com/lib/tau/detail.action?docID=5631525.
Created from tau on 2020-11-30 08:29:08.

links. Therefore, merging clusters/objects with the most number of links first during clustering process will result in more meaningful clusters.

The neighbors of a data object are those that are considerably similar to it. Measuring similarity could be done using one of the well-known distance metrics (like L_1, L_2) or even non-metric (a distance/similarity function provided by a domain expert).

Consider a data set $D = \{X_1, X_2, \ldots, X_n\}$ consisting of n objects. Let $link(X_i, X_j)$ represent the number of common neighbors between two objects X_i and X_j. Since the idea is to have each cluster with a high degree of connectivity, it is required to maximize the sum of $link(X_q, X_r)$ for data object pairs (X_q, X_r) belonging to a single cluster and at the same time minimize the sum of $link(X_q, X_s)$ for (X_q, X_s) in different clusters. This leads to the following criterion function to be maximized over k clusters.

$$E_t = \sum_{i=1}^{k} n_i * \sum_{X_q, X_r \in C_i} \frac{link(X_q, X_r)}{n_i^{1+2f(\theta)}} \tag{5.1}$$

Here, C_i denotes cluster i having n_i data objects. It is to be noted that the total number of links involving a pair of objects in cluster C_i is divided by the expected total number of links in C_i, to prevent all data objects getting assigned to a single cluster.

Similarly, for a pair of clusters (C_i, C_j), the number of cross links between them $link[C_i, C_j]$ is determined using the expression given below.

$$link[C_i, C_j] = \sum_{p_q \in C_i, p_r \in C_j} link(X_q, X_r) \tag{5.2}$$

Then, the goodness measure $g(C_i, C_j)$ for merging clusters C_i, C_j is computed as

$$g(C_i, C_j) = \frac{link[C_i, C_j]}{(n_i + n_j)^{1+2f(\theta)} - n_i^{1+2f(\theta)} - n_j^{1+2f(\theta)}} \tag{5.3}$$

The denominator in the above equation represents the expected number of cross links or links between pairs of objects each from a different cluster. The pair of clusters for which the goodness measure is maximum is the best pair to be merged at any given step.

According to this method, the larger the number of links between a pair objects, the greater is the likelihood that they belong to the same cluster. Thus, clustering using links injects global knowledge into the clustering process and is thus more robust. This is in contrast to the similarity-based local approach taking into account the characteristics of the data objects in pairs.

ROCK is a hierarchical clustering algorithm that accepts as input the set S of n sampled data objects to be clustered (that are drawn randomly from the original data

Copyright © 2019. Springer International Publishing AG. All rights reserved.

Ranga, Suri, N. N. R., et al. Outlier Detection: Techniques and Applications, Springer International Publishing AG, 2019.
ProQuest Ebook Central, http://ebookcentral.proquest.com/lib/tau/detail.action?docID=5631525.
Created from tau on 2020-11-30 08:29:08.

set), and the desired number of clusters, k. Subsequently, the clusters obtained with the sampled data are used to assign the rest of the data objects to relevant clusters.

ROCK algorithm has a worst-case time complexity of $O(n^2 + nm_m m_a + n^2 log(n))$. Here, m_m is the maximum number of neighbors, m_a is the average number of neighbors and n is the number of data objects. Similarly, the space complexity of this algorithm is $O(min\{n^2, nm_m m_a\})$.

5.2.2 Squeezer Algorithm

Squeezer algorithm performs clustering of categorical data by reading the data objects one by one. It assigns the first object to the initial cluster. Subsequently, further objects are either put into existing clusters or assigned to a new cluster depending on their similarity to the existing clusters. As a result, it doesn't require the number of desired clusters as an input parameter.

Let $D = \{X_1, X_2, \ldots, X_n\}$ be a data set having n data objects described using m categorical attributes $\{A_1, A_2, \ldots, A_m\}$. Given a cluster C_j, the set of values of the attribute A_i with respect to C_j is denoted as $VAL_i(C_j)$. Also, the support of a value a_i of the attribute A_i with respect to cluster C_j, denoted as $Sup(a_i)$, is the number of objects in C_j having the value a_i corresponding to the attribute A_i.

This algorithm determines the summary of every cluster consisting of m element pairs of attribute values and their corresponding supports.

$$Sum(C_j) = \{VS_i | 1 \leq i \leq m\} \text{where } VS_i = \{(a_i, Sup(a_i)) | a_i \in VAL_i(C_j)\}. \quad (5.4)$$

Summary of a cluster is the data structure used to compute the similarity between a data object and a cluster. A similarity threshold s is used to determine whether a data object is to be put into an existing cluster or assigned to a new cluster.

In the worst case, Squeezer algorithm has time complexity of $O(nkpm)$ and space complexity of $O(n + kpm)$. Here, k is the final number of clusters formed and p is the distinct values of a categorical attribute. It is assumed that every attribute has the same number of distinct attribute values, for simplification.

This algorithm is highly efficient for disk resident data sets as it makes only one scan over the data set for clustering the entire data. So, it is suitable for clustering data streams.

5.2.3 k-ANMI Algorithm

The k-ANMI algorithm is a k-means like clustering algorithm for categorical data. It directly optimizes the mutual information sharing based objective function. The goodness of clustering in each step is evaluated using Average Normalized Mutual Information (ANMI) measure, borrowed from cluster ensemble.

Copyright © 2019. Springer International Publishing AG. All rights reserved.

Ranga, Suri, N. N. R., et al. Outlier Detection: Techniques and Applications, Springer International Publishing AG, 2019.
ProQuest Ebook Central, http://ebookcentral.proquest.com/lib/tau/detail.action?docID=5631525.
Created from tau on 2020-11-30 08:29:08.

In information theory, mutual information is a symmetric measure to quantify the statistical information shared between two distributions. Let A and B be the random variables described by the cluster labeling $\lambda^{(a)}$ and $\lambda^{(b)}$, with $k^{(a)}$ and $k^{(b)}$ groups respectively. Let $I(A, B)$ denote the mutual information between A and B, and $H(A)$ denote the entropy of A. The [0, 1]-normalized mutual information between A and B is given by

$$NMI(A, B) = \frac{2I(A, B)}{H(A) + H(B)} \tag{5.5}$$

Let $n^{(h)}$ be the number of objects in cluster C_h according to $\lambda^{(a)}$, and let n_g be the number of objects in cluster C_g according to $\lambda^{(b)}$. Let $n_g^{(h)}$ be the number of objects in cluster C_h according to $\lambda^{(a)}$ as well as in cluster C_g according to $\lambda^{(b)}$. The [0, 1]-normalized mutual information criteria $\phi^{(NMI)}$ is computed as follows.

$$\phi^{(NMI)}(\lambda^{(a)}, \lambda^{(b)}) = \frac{2}{n} \sum_{h=1}^{k^{(a)}} \sum_{g=1}^{k^{(b)}} n_g^{(h)} log_{k^{(a)}k^{(b)}} \frac{n_g^{(h)} n}{n^{(h)} n_g} \tag{5.6}$$

Therefore, the average normalized mutual information (ANMI) between a set of m labellings, Λ, and a labeling $\bar{\lambda}$ is defined as follows.

$$\phi^{(NMI)}(\Lambda, \bar{\lambda}) = \frac{1}{m} \sum_{q=1}^{m} \phi^{(NMI)}(\bar{\lambda}, \lambda^{(q)}) \tag{5.7}$$

The optimal combined clustering $\lambda^{(k-opt)}$ will be the one that has the maximal average mutual information with all individual partitioning $\lambda^{(q)}$ given that the number of consensus clusters desired is k. Therefore, the optimal clustering is computed as

$$\lambda^{(k-opt)} = arg\ max_{\bar{\lambda}} \sum_{q=1}^{m} \phi^{(NMI)}(\bar{\lambda}, \lambda^{(q)}) \tag{5.8}$$

where $\bar{\lambda}$ goes through all possible k-partitions.

k-ANMI algorithm takes the number of desired clusters (say k) as input and iteratively changes the cluster label of each data object to improve the value of the objective function. For each object, the current label is changed to each of the other $(k - 1)$ possible labels and the ANMI objective is re-computed. If the ANMI value increases, the object's label is changed to the best new value and the algorithm proceeds with next object. A sweep over the data gets completed when all objects are checked for possible label changes. If at least one label change happens in the current sweep, a new sweep is initiated. The algorithm terminates when a local optimum is reached for the objective function.

This algorithm is easy to implement, requiring multiple hash tables as the only major data structure. More precisely, for a clustering task with m attributes and k

Copyright © 2019. Springer International Publishing AG. All rights reserved.

Ranga, Suri, N. N. R., et al. Outlier Detection: Techniques and Applications, Springer International Publishing AG, 2019.
ProQuest Ebook Central, http://ebookcentral.proquest.com/lib/tau/detail.action?docID=5631525.
Created from tau on 2020-11-30 08:29:08.

clusters, it needs $(k+1)m$ has tables. Through the use of these hash tables, ANMI value could be derived without accessing the original data set, for computational efficiency.

This algorithm has time complexity $O(tnk^2mp^2)$ in worst case. Here, t is the number of iterations the algorithm runs and p is the number of values each attribute has. Similarly, the space complexity is $O(mkp + nm)$. It is important to note that the computational complexity of k-ANMI algorithm is linear in both the number of objects (n) and the number of attributes (m). So, it can be employed for clustering large categorical data sets.

5.2.4 k-modes Algorithm

The k-modes algorithm extends the k-means paradigm to categorical domain by employing a suitable dissimilarity measure defined over categorical attributes. It replaces means of clusters with modes, and uses a frequency based method to update modes in the clustering process to minimize the associated cost function.

Compared to k-means method, there are three major modifications incorporated with k-modes method to make it work over categorical data.

1. *Dissimilarity measure*: Let X, Y be two data objects described using m categorical attributes $\{A_1, A_2, \ldots, A_m\}$. The dissimilarity between X and Y can be measured by counting the total mis-matches of the corresponding attribute categories as given in Eq. 4.1. Alternatively, dissimilarity can also be measured by taking into account the frequencies of categories in the data set as follows.

$$d_{\chi^2}(X, Y) = \sum_{j=1}^{m} \frac{(n_{x_j} + n_{y_j})}{n_{x_j} n_{y_j}} \delta(x_j, y_j) \tag{5.9}$$

   Here, n_{x_j}, n_{y_j} are the number of objects in the data set that have categories x_j and y_j for the attribute A_j. Note that $d_{\chi^2}(X, Y)$ is called as *chi-square distance*. It gives more importance to rare categories than frequent ones. So, it is useful in finding out the under-represented object clusters such as fraudulent claims in insurance databases.
2. *Mode of a set*: Let $D = \{X_1, X_2, \ldots, X_n\}$ be a data set of n data objects described using m categorical attributes. A mode of D is a vector $Z = \{z_1, z_2, \ldots, z_m\}$ that minimizes the following.

$$dist(D, Z) = \sum_{i=1}^{n} d(X_i, Z) \tag{5.10}$$

   Here, d is the dissimilarity measure as defined above or any other measure defined over categorical attributes. It is important to note that Z is not necessarily an object in D and also the mode of a data set is not unique.

Copyright © 2019. Springer International Publishing AG. All rights reserved.

Ranga, Suri, N. N. R., et al. Outlier Detection: Techniques and Applications, Springer International Publishing AG, 2019.
ProQuest Ebook Central, http://ebookcentral.proquest.com/lib/tau/detail.action?docID=5631525.
Created from tau on 2020-11-30 08:29:08.

3. Use of frequency-based method to update the cluster modes iteratively.

Let $\{S_1, S_2, \ldots, S_k\}$ be a partition of the data set, where $S_l \neq \phi$ for $1 \leq l \leq k$, and $\{Z_1, Z_2, \ldots, Z_k\}$ be the modes of $\{S_1, S_2, \ldots, S_k\}$ respectively. Then, the overall cost of the partition is given by

$$E = \sum_{l=1}^{k} \sum_{i=1}^{n} y_{i,l} d(X_i, Z_l) \tag{5.11}$$

Here, $y_{i,l}$ is an element of the partition matrix $\mathbf{Y}_{nxl}$ and d is the dissimilarity function.

Based on the three building blocks detailed above, the k-modes algorithm consists of the following steps for performing clustering of categorical data.

1. Select k initial modes, one for each cluster.
2. Allocate an object to the cluster whose mode is the nearest to it according to d. Update the mode of the cluster after each allocation.
3. After all objects are allocated to clusters, re-compute the dissimilarity of objects against the current modes. If an object is found such that its nearest mode belongs to another cluster rather than its current one, re-allocate the object to that cluster and update the mode of both clusters.
4. Repeat step-3 until no object has changed clusters over the entire data.

The initial modes of the algorithm can be selected either by considering the first k distinct objects from the data set or using any other more involved process. The advantage with this algorithm is its scalability to large data sets.

5.2.4.1 Cluster Initialization Method

A suitable cluster initialization ensures smooth convergence of the k-means family of clustering algorithms. Accordingly, a well known cluster initialization method for categorical data is discussed here. It results in a set of k clusters $\{C_1, C_2, \ldots, C_k\}$ with their representatives (modes) denoted as $\{Z_1, Z_2, \ldots, Z_k\}$.

This initialization method starts with the data object with maximum density (see Eq. 4.7) as the first cluster representative. For selecting the remaining $(k-1)$ elements of the initial cluster representatives set Z, both the density and the distance between objects (see Eq. 4.1) are utilized as per the steps furnished below.

1. Set the first cluster representative, $Z = \{X_i^1\}$, where $X_i^1 \in D$ is the object with the maximum density.
2. For the second cluster representative, $Z = Z \cup \{X_i^2\}$, where $X_i^2 \in D$ satisfies $d(X_i^2, X_m) * density(X_i^2) = max_{i=1}^{|D|}\{d(X_i, X_m) * density(X_i) | X_m \in Z\}$.
3. Similarly, for the kth cluster representative, $Z = Z \cup \{X_i^k\}$, where $X_i^k \in D$ satisfies $d(X_i^k, X_m) * density(X_i^k) = max_{i=1}^{|D|}\{min_{X_m}\{d(X_i, X_m) * density(X_i)\} | X_i \in D\}$.

Copyright © 2019. Springer International Publishing AG. All rights reserved.

Ranga, Suri, N. N. R., et al. Outlier Detection: Techniques and Applications, Springer International Publishing AG, 2019.
ProQuest Ebook Central, http://ebookcentral.proquest.com/lib/tau/detail.action?docID=5631525.
Created from tau on 2020-11-30 08:29:08.

5.3 A Ranking-Based Characterization of Outliers

Characterizing the deviating nature of data objects in categorical space is paramount to evolve accurate and efficient algorithms for their detection.

Going by the general practice, it is more meaningful to rank the data objects based on their degree of deviation rather than making a binary decision on whether or not an object is an outlier. Also, in many application domains dealing with large data, it is pertinent to identify the set of most likely outliers, as it gives opportunity for carrying out further analysis on the ranked objects. With this view, the characterization presented here leverages ranking concept for determining the set of most likely outliers in a given data set.

5.3.1 Exploring the Categorical Space

It is known that the performance of any outlier detection algorithm directly depends on the way outliers are perceived. So, a definition for outliers in terms of the attribute value frequencies of the categorical data is considered here. This definition relies on the intuition that an outlier has to display its deviating characteristics from the rest of the data in terms of irregular/peculiar value(s) corresponding to one or more of its attributes. Naturally, the frequency count of such an irregular value happens to be much less than that of the regular values. Thus, infrequent attribute values are regarded as manifestations of the presence of outliers in data.

As per the above approach, a data object X in a categorical data set D turns out to be an *outlier* in the following manner.

- *Type-1 Outlier*: The attribute values describing object X are relatively infrequent.
- *Type-2 Outlier*: The combination of the categorical values describing object X is relatively infrequent, though each one of these values are frequent individually.

These scenarios can be illustrated in a hypothetical manner as shown in Fig. 5.1, for a simple data set described using two categorical attributes. Here, data objects are depicted in categorical space based on the distinct values each attribute has in the data. Referring to this figure, object O_1 turns out to be an outlier of *Type-1* as its value corresponding to Attribute-2 is infrequent. On the other hand, though both the attribute values of object O_2 are frequent individually, their combination is not frequent, making it an outlier of *Type-2*. For object O_3, though it qualifies to be of *Type-2*, it is primarily of *Type-1* due to its infrequent attribute values. Hence, no distinction is made between objects O_1 and O_3 in this methodology.

The methods designed based on frequencies of attribute values are good at detecting outliers of Type-1. However, they fail to detect Type-2 outliers due to their characterizing feature. On the other hand, clustering-based methods are capable of detecting outliers of Type-2, but may fail in dealing with Type-1 as these outliers

Copyright © 2019. Springer International Publishing AG. All rights reserved.

Ranga, Suri, N. N. R., et al. Outlier Detection: Techniques and Applications, Springer International Publishing AG, 2019.
ProQuest Ebook Central, http://ebookcentral.proquest.com/lib/tau/detail.action?docID=5631525.
Created from tau on 2020-11-30 08:29:08.

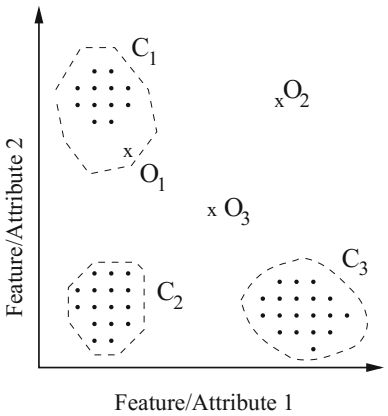


Fig. 5.1 Scenarios of outlier occurrence in categorical space

tend to be part of a valid big cluster due to their proximity to such clusters. For example, object O_1 in Fig. 5.1 may become part of its nearest cluster when clustering is performed. Ideally, an algorithm for outlier detection is expected to identify outliers of both types.

5.3.2 Methodology for Characterization

Consider a data set D comprising n objects described using m categorical attributes. The basic idea is to capture all possible types of outliers considering their occurrence scenarios in the categorical space. Also, such a characterization is to be done based on the frequency counts of the values of categorical attributes describing the data. In this connection, the definition of outliers given above looks more appealing.

In order to determine the most likely set of outliers in the data, a clustering-based methodology is presented here. For the sake of illustration, let us assume a clustering structure of some hypothetical categorical data as shown in Fig. 5.2. To start with, some important definitions constituting this characterization are given below.

1. A cluster C_i is considered as a *big cluster* if it has at least $\alpha\%$ of the number of objects in D. Thus, the set of big clusters BC in a clustering process is given by

$$BC = \{C_i | size(C_i) \geq (\alpha * 0.01 * n)\} \tag{5.12}$$

   where $size(C_i)$ indicates the number objects present in cluster C_i and value of α is determined specific to the data set at hand.
2. The *cluster distance* of an object X_i is measured as the distance between X_i and its closest big cluster as,

$$cdist(X_i) = min_l\, d(Z_l, X_i), \forall C_l \in BC \tag{5.13}$$

Copyright © 2019. Springer International Publishing AG. All rights reserved.

Ranga, Suri, N. N. R., et al. Outlier Detection: Techniques and Applications, Springer International Publishing AG, 2019.
ProQuest Ebook Central, http://ebookcentral.proquest.com/lib/tau/detail.action?docID=5631525.
Created from tau on 2020-11-30 08:29:08.

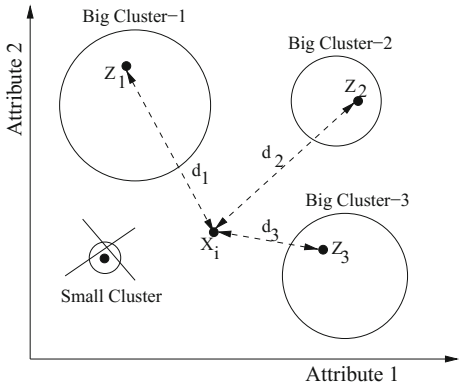


Fig. 5.2 Illustration of cluster distance computation

where Z_l is the representative of cluster C_l and the distance function d refers to the one given in Eq. 4.6.

The computation of cluster distance of an object X_i in a clustering scenario, involving big as well as small clusters, is depicted in Fig. 5.2. In this case, the big cluster with its mode represented by Z_3 turns out to be the nearest big cluster to object X_i, as distance d_3 happens to be the smallest of the distances $\{d_1, d_2, d_3\}$.

The following two aspects define the necessary conceptual details required to determine the most likely outliers in a data set.

1. As per Eq. 4.7, the density of a data object is measured based on the frequencies of its attributes values. One can note that the density of an outlier will be generally low, as the attribute values of such an object tend to be less frequent compared to the values describing normal ones. Thus, density can be looked at as a measure of the relative frequency of the values of a data object. The less the density is, the more likely the object gets detected as outlier. Based on this understanding, the data objects can be arranged in ascending order of their density values, with the least density object displaying the most deviating characteristics. Such a sequence of objects is called as *frequency-based ranking* of data.
2. Similarly, it is possible to identify small clusters possibly having outlier objects that tend to locate away from big clusters, as shown in Eq. 5.12. Therefore, the distance between a data object and its closest big cluster is an indicative measure of its deviating behavior. The more this distance is, the more likely the object gets detected as outlier. Based on this rationale, the *clustering-based ranking* of data objects is determined. More precisely, clustering-based ranking is obtained by arranging the objects in decreasing order of their cluster distance values.

According to the above two ranking schemes, each object $X \in D$ gets its corresponding ranks based on its density and its cluster distance values. Given that, outliers of Type-1 can be detected using the frequency-based ranking, and Type-2 using the clustering-based ranking. Now, a consolidated set of outliers of both types, designed as the *likely set* (LS), can be determined using these two ranking schemes.

Copyright © 2019. Springer International Publishing AG. All rights reserved.

Ranga, Suri, N. N. R., et al. Outlier Detection: Techniques and Applications, Springer International Publishing AG, 2019.
ProQuest Ebook Central, http://ebookcentral.proquest.com/lib/tau/detail.action?docID=5631525.
Created from tau on 2020-11-30 08:29:08.

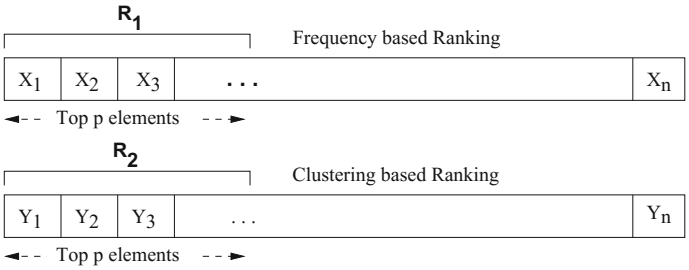


Fig. 5.3 Determining the likely set of outliers in data

To determine p most significant outliers in data, consider the set of indices of p top ranked objects in the frequency-based ranking as R_1. Similarly, the set of indices of p top ranked objects in the clustering-based ranking is denoted as R_2, as shown in Fig. 5.3. Now, the likely set of outliers is constructed as, $LS = R_1 \cup R_2$. Though there are several possible ways of fusing two rank lists, simple set union can produce the likely set satisfactorily.

5.4 Algorithms for Detecting Outliers

As stated earlier, most of the methods for outlier detection are designed to deal with data sets described by numerical attributes, or ordinal attributes that can be directly mapped into acceptable numerical values. Only a limited number of the approaches can process categorical data. This is due to the challenges posed by the categorical attributes in defining requisite measures/metrics for exploring the data. However, with emphasis on this aspect of the outlier detection problem, there are some interesting methods employing different detection strategies. Some of the well known approaches among them for outlier detection in categorical data are discussed below.

5.4.1 Greedy Algorithm

Conventional approaches for exploratory data mining applications do not handle categorical data in a satisfactory manner. To address this gap, the problem of outlier detection in categorical data is defined as an optimization problem in terms of finding a sub-set of k objects such that the expected entropy of the resultant data set after the removal of this sub-set is minimized.

However, an exhaustive search through all possible solutions with k outliers for the one with the minimum objective value is costly since there are (n, k) possible

Copyright © 2019. Springer International Publishing AG. All rights reserved.

Ranga, Suri, N. N. R., et al. Outlier Detection: Techniques and Applications, Springer International Publishing AG, 2019.
ProQuest Ebook Central, http://ebookcentral.proquest.com/lib/tau/detail.action?docID=5631525.
Created from tau on 2020-11-30 08:29:08.

solutions with n objects and k outliers. So, a fast greedy algorithm for mining outliers under the same optimization model is presented here.

For a discrete random variable X with $S(X)$ being the set of values that it can take, $p(x)$ being the probability distribution function, the entropy of X can be computed using Eq. 4.22.

Given a data set D of n objects $\{X_1, X_2, \ldots, X_n\}$, where each object is described using m categorical attributes $\{f_1, f_2, \ldots, f_m\}$, then the entropy of D can be computed as

$$E(D) = E(f_1) + E(f_2) + E(f_3) + \cdots + E(f_m) \tag{5.14}$$

Here, $E(f_j)$ indicates the entropy value (as in Eq. 4.22) corresponding to the attribute/feature f_j of the data set D. The above expression holds good with the assumption of independent attributes.

Given an integer k, the idea is to determine a subset $D^{'} \subset D$ with size k, in such a way that the entropy of the resulting set $(D - D^{'})$ is minimized. According to this idea, a data object with maximum contribution to the entropy value gets labeled as the first outlier. It goes on identifying the outliers in successive iterations at the rate of one outlier per iteration. Hence, this algorithm requires a number of scans over the data set resulting in a time complexity of $O(nkm)$.

5.4.2 AVF Algorithm

Attribute Value Frequency (AVF) algorithm is a fast and scalable outlier detection strategy for categorical data. It is developed based on the intuition that outliers are those objects that are infrequent in the data set. Moreover, an ideal outlier object in a categorical data set is one whose each and every attribute value is extremely irregular (or infrequent). The infrequent-ness of an attribute value can be determined by counting the number of times it appears in the data set.

Considering the above discussion, the AVF method works based on the frequency counts of the attribute values. Given a data set with n data objects and m categorical attributes, this method computes the frequency-based score of an object X_i as

$$AVFScore(X_i) = \frac{1}{m} \sum_{j=1}^{m} Freq(x_{ij}) \tag{5.15}$$

where $Freq(x_{ij})$ is the number of times the jth attribute/feature value of the object X_i appears in the data set.

As per this method, objects with low AVF-scores are considered as outliers. Once the scores of all the data objects are calculated, one can designate the k objects with the smallest AVF-scores as outliers.

The computational complexity of this algorithm is $O(nm)$. It scales linearly with the number of data objects and attributes, and works with a single scan over the data

Copyright © 2019. Springer International Publishing AG. All rights reserved.

Ranga, Suri, N. N. R., et al. Outlier Detection: Techniques and Applications, Springer International Publishing AG, 2019.
ProQuest Ebook Central, http://ebookcentral.proquest.com/lib/tau/detail.action?docID=5631525.
Created from tau on 2020-11-30 08:29:08.

set. It does not need to create or search through different combinations of attribute values.

5.4.3 ROAD Algorithm

As per the characterization of outliers presented in the previous section, the computational flow of a ranking-based algorithm for unsupervised detection of outliers is shown in Fig. 5.4. Basically, this algorithm performs its computations in two phases. In the first phase, it computes the object density values and also explores a clustering of the data. Using the resultant clustering structure, the set of big clusters is identified in order to determine the distance between various data objects and their corresponding nearest big clusters. In the second phase, frequency-based rank and clustering-based rank of each data object are determined. Subsequently, a unified set of the most likely outliers is determined using these two individual rankings. Thus, it is named as Ranking-based Outlier Analysis and Detection (ROAD) algorithm.

Clustering categorical data is an important computational step of the ROAD algorithm, as it is meant for identifying Type-2 outliers which are relatively more difficult to detect when compared to Type-1 outliers. This is because, Type-2 outliers display their deviation in the joint distribution of two or more attributes. Accordingly, the

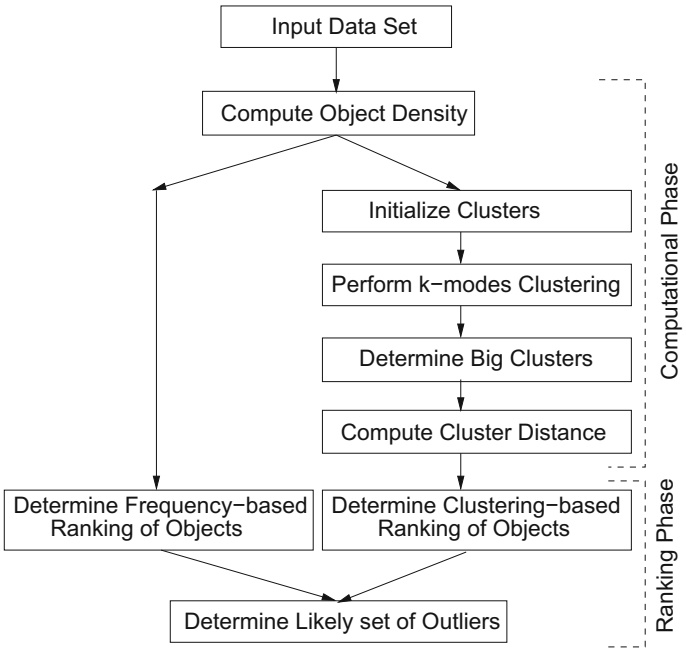


Fig. 5.4 Computational flow of ROAD algorithm

Copyright © 2019. Springer International Publishing AG. All rights reserved.

Ranga, Suri, N. N. R., et al. Outlier Detection: Techniques and Applications, Springer International Publishing AG, 2019.
ProQuest Ebook Central, http://ebookcentral.proquest.com/lib/tau/detail.action?docID=5631525.
Created from tau on 2020-11-30 08:29:08.

effectiveness of the specific clustering method employed for this purpose determines the detection performance.

In practice, one can use any established method like ROCK algorithm for clustering categorical data. However, it is recommended to use the *k-modes* method due to its efficiency as a member of the *k-means* family. Moreover, the cluster initialization method presented in Sect. 5.2.4.1 makes it more effective. Hence, using the k-modes algorithm along with the categorical dissimilarity measure given in Eq. 4.6 is expected to exhibit better results.

The data processing steps constituting the ROAD algorithm are furnished below. The initial five steps constitute the first phase known as 'computational phase' as it primarily computes various quantities on the basis of the input data. The later three steps correspond to the second phase called as 'ranking phase', which determines the rank of each data object corresponding to each ranking scheme based on the quantities computed in the previous phase.

1. Compute $density(X_i)$ of each data object $X_i \in D$ using Eq. 4.7.
2. Determine the initial set $\{Z_1, Z_2, \ldots, Z_k\}$ of k cluster representatives.
3. Perform k-modes clustering on D with the distance measure in Eq. 4.6.
4. Determine the set of big clusters BC as given in Eq. 5.12.
5. For each object X_i, determine its cluster distance $cdist(X_i)$ as per Eq. 5.13.
6. Determine the frequency-based rank $freq\_rank(X_i)$ of each object $X_i \in D$.
7. Determine the clustering-based rank $clust\_rank(X_i)$ of each object $X_i \in D$.
8. Construct the likely set LS using the two ranked sequences, for a given p value.
9. Output the the set LS of most likely outliers identified.

Considering various simplifications on the algorithmic steps, the computational complexity of ROAD algorithm turns out to be $O(nm + nlog(n))$. It is important to note that this complexity is not affected by the number of outliers to be found.

5.5 Experiments on Benchmark Data

Experimental study of the outlier detection algorithms on benchmark data includes six frequently used categorical data sets taken from UCI ML Repository. Each data set consists of labeled instances belonging to two different classes. As per the standard procedure explained in Sect. 4.2.2, objects with missing attribute values are removed and objects belonging to small sized class in every data set are considered as outliers. Though these designated outliers are not outliers in real sense, they are considered so for experimental purpose. In order to impose imbalance in the number of objects belonging to normal and outlier categories, only a selected sub-set of the objects belonging to outlier class are considered, by taking every fifth object of the designated outlier class. Table 5.4 summarizes the benchmark categorical data sets considered in this investigation.

Copyright © 2019. Springer International Publishing AG. All rights reserved.

Ranga, Suri, N. N. R., et al. Outlier Detection: Techniques and Applications, Springer International Publishing AG, 2019.
ProQuest Ebook Central, http://ebookcentral.proquest.com/lib/tau/detail.action?docID=5631525.
Created from tau on 2020-11-30 08:29:08.

Table 5.4 Details of benchmark categorical data sets

Name of the data set	Dimension	Outlier class label	# Outlier objects	# Normal objects	# Total objects
Chess (End-Game)	36	Nowin	305	1669	1974
Tic-Tac-Toe	9	Negative	66	626	692
Breast Cancer (W)	10	4 (malignant)	47	444	491
Congressional Votes	16	Republican	21	124	145
Breast Cancer	9	Recurrence-events	16	196	212
Mushroom	22	p (poisonous)	783	4208	4991

As the algorithms considered for this study work in unsupervised learning mode, they do not require labeled data. However, class labels are used to assess their performance in mining outliers. A comparative view of the performance of these methods is also presented here.

As described in Sect. 5.1, a categorical attribute A_i can take a set of discrete values (categories) represented by $DOM\,(A_i)$. The cardinality of the attribute domains of 9 categorical attributes describing the Breast Cancer data set are shown in Table 5.5 for a better understanding.

Table 5.5 Categorical attributes describing Breast Cancer data

Sl.No.	Attribute name	Arity
1	Age	9
2	Menopause	3
3	Tumor-size	12
4	Inv-nodes	13
5	Node-caps	2
6	Deg-malig	3
7	Breast	2
8	Breast-quad	5
9	Irradiat	2

Copyright © 2019. Springer International Publishing AG. All rights reserved.

Ranga, Suri, N. N. R., et al. Outlier Detection: Techniques and Applications, Springer International Publishing AG, 2019.
ProQuest Ebook Central, http://ebookcentral.proquest.com/lib/tau/detail.action?docID=5631525.
Created from tau on 2020-11-30 08:29:08.

5.5.1 Performance Measurement

To measure the detection accuracy of ROAD algorithm, the number of actual outliers (as per the designated outlier class labels) present among the elements of the likely set is considered. Following this measure, the results obtained on the benchmark data sets are reported in Table 5.6 under the column labeled 'Combined'. For a comparative view, performance of the AVF and the Greedy algorithms on the same data is indicated here. The best performance obtained corresponding to each data set is shown in bold font.

From this experimentation, one can understand that the ROAD algorithm exhibited better results over the other two methods. This is due to the additional capability of this algorithm in detecting outliers belonging to *Type-2*. To understand this better, the number of Type-1 and Type-2 outliers detected are shown separately in Table 5.6. The number in brackets under the column labeled 'Type-2' indicates the value of the parameter k using which these results are obtained. A fixed value 5 is used for the other parameter α of the algorithm.

As described in Sect. 4.5, experimental results of outlier detection algorithm are typically reported using ROC curves. For increasing values of top portions (p) of the ranked outlier list, TPR and FPR values are determined to produce corresponding ROC curves of all the three algorithms. The curves thus generated on the benchmark data sets are shown in Fig. 5.5 for a quick understanding of their overall performance.

5.5.2 Sensitivity Study

The sensitivity of ROAD algorithm is explored here with respect to its parameters, namely k and α. The impact of these parameters on the performance of the algorithm

Table 5.6 Performance comparison on benchmark data sets

Name of the data set	# Outliers present (p value)	# Outliers detected (among the top p objects)				
		ROAD algorithm			AVF algorithm	Greedy algorithm
		Type-1	Type-2	Combined		
Chess (End-Game)	305	78	115 (8)	**126**	78	90
Tic-Tac-Toe	66	17	30 (14)	**37**	17	8
Breast Cancer (W)	47	37	32 (3)	**42**	37	41
Congressional Votes	21	16	2 (2)	**18**	16	16
Breast Cancer	16	4	5 (3)	**5**	4	3
Mushroom	783	454	132 (8)	**575**	454	340

Copyright © 2019. Springer International Publishing AG. All rights reserved.

Ranga, Suri, N. N. R., et al. Outlier Detection: Techniques and Applications, Springer International Publishing AG, 2019.
ProQuest Ebook Central, http://ebookcentral.proquest.com/lib/tau/detail.action?docID=5631525.
Created from tau on 2020-11-30 08:29:08.

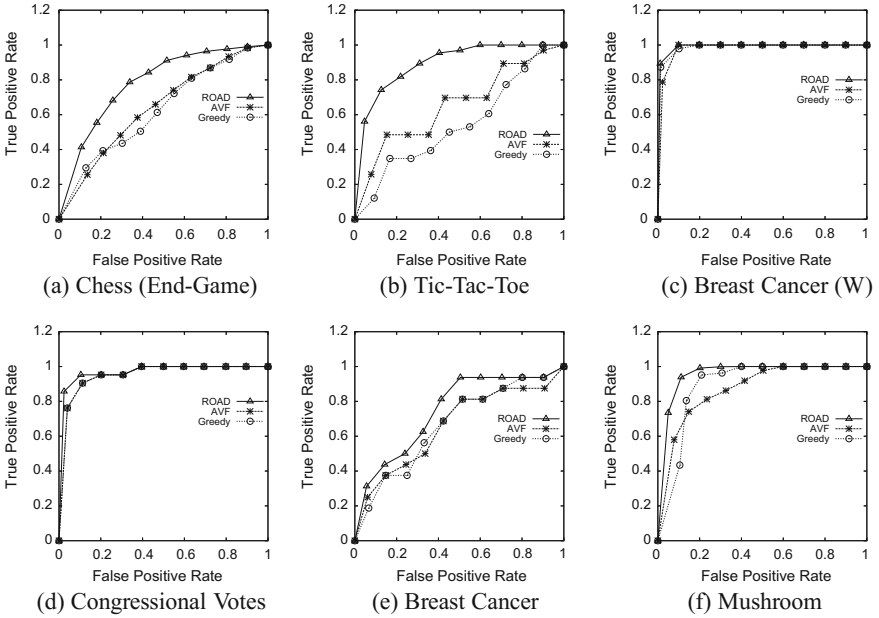


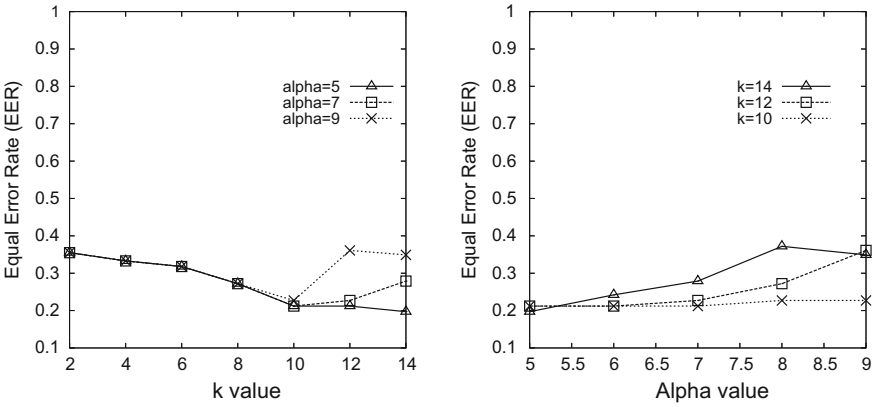
Fig. 5.5 Performance comparison using benchmark data sets

is measured using the Equal Error Rate (EER) measure, as described in Sect. 4.5. In this study, sensitivity of the algorithm is explored using Tic-Tac-Toe data shown in Table 5.4. The effect of varying the number of clusters k (keeping the value of α fixed at various values 5, 7 and 9) is as shown in Fig. 5.6a. This figure indicates that with increasing k value, the performance of ROAD algorithm has improved, as characterized by low EER values. However, arbitrarily high values of k may result in more number of small sized clusters, or the clustering step may not converge.

The observations made in a similar study with respect to parameter α, for various fixed values of k, are shown in Fig. 5.6b. According to this figure, with the increase in the value of α, the number of big clusters has reduced and this has negative impact on the performance, more so for high values of k. Moreover, with high values of α, there may not be any big clusters left out reducing the number of Type-2 outliers detected to zero. On the other hand, a very low value of α is not desirable as it cannot differentiate even the known small clusters from big ones. It is important to note that the parameters k and α are not independent. Thus, the set of allowable values for α depends on the value assigned to k.

Copyright © 2019. Springer International Publishing AG. All rights reserved.

Ranga, Suri, N. N. R., et al. Outlier Detection: Techniques and Applications, Springer International Publishing AG, 2019.
ProQuest Ebook Central, http://ebookcentral.proquest.com/lib/tau/detail.action?docID=5631525.
Created from tau on 2020-11-30 08:29:08.



(a) parameter k (b) parameter α

Fig. 5.6 Sensitivity analysis on Tic-Tac-Toe data set

5.6 Current Research Trends

Having seen numerous methods for outlier detection in categorical data, it is time to look at the current research trends based on some recent research efforts towards this problem.

5.6.1 Outlier Detection in Categorical Data Using Holoentropy

An information theoretic outlier detection method based on the concept of *weighted holoentropy* combines the entropy and the total correlation measures with attribute weighting for identifying exceptional objects. The rationale behind this method is to assign weights to the individual attributes so as to give more importance to those attributes with small entropy values for increasing the impact of removing an outlier candidate that is outstanding on those attributes. Hence, the weight computation for the attribute entropy employs a reverse sigmoid function of the entropy itself.

Based on the above description, a novel definition for outliers turns out to be the set of objects with greatest decrease in the weighted holoentropy value when deleted from the original data set. Similarly, the outlier factor computation is also modified to include weighted holoentropy by defining approximate differential holoentropy. As a result, the outlier factor (OF) of an object X_i is computed as

$$OF(X_i) = \sum_{r=1}^{m} OF(x_{i,r}) = \sum_{r=1}^{m} \begin{cases} 0, & \text{if } freq(x_{i,r}) = 1, \\ w_{\psi}(A_i).\delta[freq(x_{i,r})], & \text{else.} \end{cases} \quad (5.16)$$

Copyright © 2019. Springer International Publishing AG. All rights reserved.

Ranga, Suri, N. N. R., et al. Outlier Detection: Techniques and Applications, Springer International Publishing AG, 2019.
ProQuest Ebook Central, http://ebookcentral.proquest.com/lib/tau/detail.action?docID=5631525.
Created from tau on 2020-11-30 08:29:08.

This method, named as Information Theory Based Step by Step (ITB-SS), determines an anomaly candidate set (AS) and an upper bound on the number of likely outliers (UO). In order to determine the final set of outliers, only the objects in the AS set are examined thereby reducing the time complexity of the algorithm to $O(om(UO))$, where o is the number of outliers to be identified and m is the data dimensionality. Through an empirical study, it is found that the average UO is about $0.21n$, where n the number of objects in the data. However, the number of objects UO to be examined increases significantly for large data sets.

5.6.2 Detecting Outliers in Categorical Data Streams

Most of the techniques detecting anomalies in data streams are either distance-based or cluster-based. Both these varieties require distance computation to achieve their intended purpose. Due to the known limitations of categorical attributes, these methods lack the ability to handle such data in a proper manner. To effectively solve this problem, the concept of mining closed frequent patterns is employed. Based on this idea, the method designed for identifying outliers in categorical data streams is named as Detecting Outliers in Sliding Window over Categorical Data Streams (DOSW_CDStream).

In the context of data streams, *support* of an item set X turns out to be the ratio of the number of categorical data that contain X to the width of the sliding window w. An item set X can be considered as *closed frequent pattern* over the current sliding window, if its support is greater than the minimum support and there is no super-set of X with same support value.

The width of the sliding window is determined by two sliding end points. In this model, old data objects are thrown out as the new data arrive. The current sliding window consisting of w data objects is denoted as

$$SW_{n-w+1} = \{T_{n-w+1}, T_{n-w+2}, \ldots, T_n\} \tag{5.17}$$

Here, n indicates the current time. In order to determine outliers, this method introduced a formula named as Weighted Closed Frequent Pattern Outlier Factor (WCFPOF). For an arbitrary categorical data object T_i, its outlier factor is measured as

$$WCFPOF(T_i) = \frac{\sum_{P \subseteq T_i, P \in CFPS(DS, minsup)} sup(P) * (|P|/|T_i|)}{|CFPS(DS, minsup)|} \tag{5.18}$$

where, $CFPS(DS, minsup)$ is the set of closed frequent patterns in the current window, $|P|$ is the length of closed frequent pattern P and $minsup$ is the minimum support specified by the user. A smaller value of this outlier factor indicates a greater possibility of the corresponding object being an outlier. This method first determines all the closed patterns and then computes their outlier factors. The summary details are

Copyright © 2019. Springer International Publishing AG. All rights reserved.

Ranga, Suri, N. N. R., et al. Outlier Detection: Techniques and Applications, Springer International Publishing AG, 2019.
ProQuest Ebook Central, http://ebookcentral.proquest.com/lib/tau/detail.action?docID=5631525.
Created from tau on 2020-11-30 08:29:08.

stored in an index structure, which can be updated by a decay function. Finally, top-k outliers are determined based on a ranking over the outlier factor values. The computational cost of this method is noted as $O(CFPS + w * N + 2w + (2w * log(2w)))$.

5.7 Summary

An important research issue concerning the outlier detection problem, namely dealing with data described using categorical attributes is studied here. A majority of the outlier detection methods mainly deal with numerical data. However, data pertaining to several real life applications tend to be categorical in nature with high dimensionality. So, it is necessary to have relevant techniques to process such data.

Addressing this requirement, some important research efforts that have gone in this direction are discussed in this chapter.

- Issues and challenges in dealing with categorical data analysis are highlighted.
- Emphasizing on the role of clustering as an unsupervised learning task, a few algorithms for clustering categorical data are presented.
- A ranking-based characterization of outliers in categorical space is presented along with the associated methodology.
- A sample set of algorithms for detecting outliers in categorical data are presented along with a discussion on the detection strategies employed by each method.
- Experimental study using benchmark categorical data sets is presented to facilitate a practical view of the outlier detection problem.

Towards the end, a brief discussion on a few current research trends regarding the outlier detection problem is presented.

5.8 Bibliographic Notes

Outlier detection problem has numerous applications such as fraud detection in financial transactions and intrusion detection in computer networks, etc [6, 18]. Various research issues are involved in outlier detection and so various methods were developed employing varied detection strategies [7, 17, 23, 27, 30, 33]. Clustering-based methods [21, 29] try to identify various groups of objects based on their intrinsic similarity in order to isolate objects with deviating characteristics. The LOF method [4] is a frequently used one for detecting local outliers in numerical data.

As stated at the beginning, the theme of this chapter is to delve on the methods dealing with categorical attributes. As brought out in [3], the characteristics of a categorical data set can be summarized using a simplified representation. The main computational challenge in this regard is that various values taken by a categorical variable are not inherently ordered [2]. Thus, evolving necessary methods to deal with categorical data is a non-trivial task.

Copyright © 2019. Springer International Publishing AG. All rights reserved.

Ranga, Suri, N. N. R., et al. Outlier Detection: Techniques and Applications, Springer International Publishing AG, 2019.
ProQuest Ebook Central, http://ebookcentral.proquest.com/lib/tau/detail.action?docID=5631525.
Created from tau on 2020-11-30 08:29:08.

Similar to LOF, a method named as Cluster Based Local Outlier Factor (CBLOF) was proposed [15] for detecting local outliers in categorical data. Subsequently, an approach of comparing against marginal distributions of attribute subsets was proposed [9] in an effort to identify anomalies in categorical data in an unsupervised manner. Another such effort is the Greedy algorithm [14], based on the concept of entropy of a data set. In order to detect p outliers in a data set, this algorithm requires p scans over the data, which is computationally expensive for large data sets. Addressing this issue, the AVF algorithm [22] was proposed using the attribute value frequencies. Each data object is assigned a frequency score and objects with low scores are considered as outliers. Though this algorithm can identify objects with infrequent attribute values as outliers, it doesn't explore the case of outliers arising due to infrequent combination of attributes with frequent values.

A distance-based method for outlier detection in high-dimensional categorical data was proposed in [23] using the notion of common neighbor of a pair of objects. According to this method, the distance between two objects X_i and X_j is defined as

$$dist(X_i, X_j, \theta) = 1 - \log_2 |CNS(X_i, X_j, \theta)| / \log_2 |D|$$

where

$$CNS(X_i, X_j, \theta) = NS(X_i, \theta) \cap NS(X_j, \theta).$$

Here, θ is a user defined threshold, $NS(X_i, \theta)$ is the set of neighbors of X_i. Though this method captures dissimilarity between objects driven by the global data distribution, its time complexity turns out to be quadratic in the number of objects. A genetic approach for outlier detection in projected space was proposed [1], which is applicable for both numeric and categorical data. Another recent effort is the SCF (Squares of the Complement of the Frequency) algorithm proposed in [29] using the attribute value frequencies like the AVF algorithm.

Given the unsupervised nature of the outlier detection problem, clustering-based methods are natural choices. Robust Clustering using linKs (ROCK) algorithm [13] works based on the notion of links between data objects, rather than applying any metric to measure the similarity. Squeezer algorithm [32] performs clustering of categorical data by reading the data objects one by one. The k-ANMI algorithm [16] is a k-means like clustering algorithm that directly optimizes the mutual information sharing based objective function.

It is known that the performance of any outlier detection algorithm directly depends on the way outliers are perceived [8]. Therefore, a customized definition meeting the specific requirements of an application is necessary to proceed with the detection of outliers. As brought out in [24], it is more meaningful to rank the data objects based on their degree of deviation. Accordingly, a ranking-based formalism, named as ROAD algorithm, for outlier detection in categorical data was presented in [26, 28] using data clustering as the basis. The objective was to capture all possible types of outliers considering their occurrence scenarios in the categorical space. In this connection, the k-modes algorithm [19] was employed for clustering due to its efficiency as a member of the *k-means* family [20] of algorithms. Moreover,

Copyright © 2019. Springer International Publishing AG. All rights reserved.

Ranga, Suri, N. N. R., et al. Outlier Detection: Techniques and Applications, Springer International Publishing AG, 2019.
ProQuest Ebook Central, http://ebookcentral.proquest.com/lib/tau/detail.action?docID=5631525.
Created from tau on 2020-11-30 08:29:08.

employing the cluster initialization method [5] along with the categorical dissimilarity measure [25] made it more effective.

Experimental study of these methods can be carried out on various benchmark categorical data sets available at UCI ML Repository [10]. Typically, results of outlier detection algorithm are reported using ROC curves [12]. Likewise, the impact of the algorithmic parameters on its performance can be measured using the Equal Error Rate (EER) measure [21].

Some of the recent methods for dealing with categorical data are also presented in this chapter. An information theoretic outlier detection algorithm was proposed recently [31] by defining the concept of weighted holoentropy. Similarly, an algorithm for identifying outliers in categorical data streams was formulated in [30] based on the concept of mining closed frequent patterns.

References

1. Bandhyopadhyay, S., Santra, S.: A genetic approach for efficient outlier detection in projected space. Pattern Recognit. **41**, 1338–1349 (2008)
2. Bock, H.H.: The classical data situation. In: Analysis of Symbolic Data, pp. 139–152. Springer (2002)
3. Boriah, S., Chandola, V., Kumar, V.: Similarity measures for categorical data: a comparative evaluation. In: SIAM International Conference on Data Mining, Atlanta, Georgia, USA, pp. 243–254 (2008)
4. Breunig, M., Kriegel, H., Ng, R., Sander, J.: Lof: identifying density-based local outliers. In: ACM SIGMOD International Conference on Management of Data, Dallas, Texas, pp. 93–104 (2000)
5. Cao, F., Liang, J., Bai, L.: A new initialization method for categorical data clustering. Expert. Syst. Appl. **36**, 10223–10228 (2009)
6. Chandola, V., Banerjee, A., Kumar, V.: Anomaly detection: a survey. ACM Comput. Surv. **41**(3) (2009)
7. Chandola, V., Banerjee, A., Kumar, V.: Anomaly detection for discrete sequences: a survey. IEEE Trans. Knowl. Data Eng. (TKDE) **24**(5), 823–839 (2012)
8. Cui, Z., Ramanna, S., Peters, J.F., Pal, S.K.: Cognitive informatics and computational intelligence: theory and applications. Fundam. Inform. **124**(1–2), v–viii (2013)
9. Das, K., Schneider, J.: Detecting anomalous records in categorical datasets, San Jose, California. In: ACM KDD, pp. 220–229 (2007)
10. Dua, D., Efi, K.T.: UCI machine learning repository (2017). http://archive.ics.uci.edu/ml
11. Duan, L., Xu, L., Liu, Y., Lee, J.: Cluster-based outlier detection. Ann. Oper. Res. **168**, 151–168 (2009)
12. Fawcett, T.: An introduction to ROC analysis. Pattern Recognit. Lett. **27**, 861–874 (2006)
13. Guha, S., Rastogi, R., Kyuseok, S.: ROCK: A robust clustering algorithm for categorical attributes. In: International Conference on Data Engineering (ICDE), Sydney, Australia, pp. 512–521 (1999)
14. He, Z., Xu, X., Deng, S.: A fast greedy algorithm for outlier mining. In: Proceedings of Pacific Asia Conference on Knowledge Discovery in Databases (PAKDD), Singapore, pp. 567–576 (2006)
15. He, Z., Xu, X., Deng, S.: Discovering cluster-based local outliers. Pattern Recognit. Lett. **24**, 1641–1650 (2003)
16. He, Z., Xu, X., Deng, S.: k-ANMI: a mutual information based clustering algorithm for categorical data. Inf. Fusion **9**, 223–233 (2008)

Copyright © 2019. Springer International Publishing AG. All rights reserved.

Ranga, Suri, N. N. R., et al. Outlier Detection: Techniques and Applications, Springer International Publishing AG, 2019.
ProQuest Ebook Central, http://ebookcentral.proquest.com/lib/tau/detail.action?docID=5631525.
Created from tau on 2020-11-30 08:29:08.

17. Hido, S., Tsuboi, Y., Kashima, H., Sugiyama, M., Kanamori, T.: Statistical outlier detection using direct density ratio estimation. Knowl. Inf. Syst. **26**(2), 309–336 (2011)
18. Hodge, V.J., Austin, J.: A survey of outlier detection methodologies. Artif. Intell. Rev. **22**, 85–126 (2004)
19. Huang, Z.: A fast clustering algorithm to cluster very large categorical data sets in data mining. In: SIGMOD Data Mining and Knowledge Discovery Workshop, pp. 1–8 (1997)
20. Jain, A.K.: Data clustering: 50 years beyond K-means. Pattern Recognit. Lett. **31**, 651–666 (2010)
21. Jain, A.K., Duin, R.P.W., Mao, J.: Statistical pattern recognition: a review. IEEE Trans. Pattern Anal. Mach. Intell. **22**, 4–37 (2000)
22. Koufakou, A., Ortiz, E., Georgiopoulos, M.: A scalable and efficient outlier detection strategy for categorical data. In: Proceedings of IEEE ICTAI, Patras, Greece, pp. 210–217 (2007)
23. Li, S., Lee, R., Lang, S.D.: Mining distance-based outliers from categorical data. In: IEEE ICDM Workshop, Omaha, Nebraska, pp. 225–230 (2007)
24. Muller, E., Assent, I., Steinhausen, U., Seidl, T.: Outrank: ranking outliers in high dimensional data. In: IEEE ICDE Workshop, Cancun, Mexico, pp. 600–603 (2008)
25. Ng, M.K., Li, M.J., Huang, J.Z., He, Z.: On the impact of dissimilarity measure in k-modes clustering algorithm. IEEE Trans. Pattern Anal. Mach. Intell. **29**(3), 503–507 (2007)
26. Suri, N.N.R.R., Murty, M., Athithan, G.: An algorithm for mining outliers in categorical data through ranking. In: 12th International Conference on Hybrid Intelligent Systems (HIS), pp. 247–252. IEEE Xplore, Pune, India (2012)
27. Suri, N.N.R.R., Murty, M., Athithan, G.: Data mining techniques for outlier detection. In: Zhang, Q., Segall, R.S., Cao, M. (eds.) Visual Analytics and Interactive Technologies: Data, Text and Web Mining Applications, Chap. 2, pp. 22–38. IGI Global, New York, USA (2011)
28. Suri, N.N.R.R., Murty, M., Athithan, G.: A ranking-based algorithm for detection of outliers in categorical data. Int. J. Hybrid Intell. Syst. (IJHIS) **11**(1), 1–11 (2014)
29. Taha, A., Hegazy, O.M.: A proposed outliers identification algorithm for categorical data sets. In: 7th International Conference on Informatics and Systems (INFOS), Cairo, Egypt, pp. 1–5 (2010)
30. Wu, Q., Ma, S.: Detecting outliers in sliding window over categorical data streams. In: 8th International Conference on Fuzzy Systems and Knowledge Discovery (FSKD), pp. 1663–1667. IEEE (2011)
31. Wu, S., Wang, S.: Information-theoretic outlier detection for large-scale categorical data. IEEE Trans. Knowl. Data Eng. (TKDE) **25**(3), 589–602 (2013)
32. Zengyou, H., Xiaofei, X., Shengchun, D.: Squeezer: an efficient algorithm for clustering categorical data. J. Comput. Sci. Technol. **17**(5), 611–624 (2002)
33. Zhang, K., Hutter, M., Jin, H.: A new local distance-based outlier detection approach for scattered real-world data. In: PAKDD, Bangkok, Thailand, pp. 813–822 (2009)

Copyright © 2019. Springer International Publishing AG. All rights reserved.

Ranga, Suri, N. N. R., et al. Outlier Detection: Techniques and Applications, Springer International Publishing AG, 2019.
ProQuest Ebook Central, http://ebookcentral.proquest.com/lib/tau/detail.action?docID=5631525.
Created from tau on 2020-11-30 08:29:08.