

A Framework for Exploring Categorical Data

Varun Chandola Shyam Boriah Vipin Kumar

Department of Computer Science and Engineering
University of Minnesota

<chandola,sboriah,kumar>@cs.umn.edu

Abstract

In this paper, we present a framework for categorical data analysis which allows such data sets to be explored using a rich set of techniques that are only applicable to continuous data sets. We introduce the concept of separability statistics in the context of exploratory categorical data analysis. We show how these statistics can be used as a way to map categorical data to continuous space given a labeled reference data set. This mapping enables visualization of categorical data using techniques that are applicable to continuous data. We show that in the transformed continuous space, the performance of the standard k -nn based outlier detection technique is comparable to the performance of the k -nn based outlier detection technique using the best of the similarity measures designed for categorical data. The proposed framework can also be used to devise similarity measures best suited for a particular type of data set.

1 Introduction

Categorical data (also known as nominal or qualitative multi-state data) has become increasingly common in modern real-world applications. Table 1 shows a sample of a categorical data set. These data sets are often rich in information and are frequently encountered in domains where large-scale data sets are common, e.g., in network intrusion detection. However, unlike continuous data, categorical data attribute values cannot be naturally mapped on to a scale, making most continuous data analysis techniques inapplicable in this setting: Table 2 lists common exploratory analysis techniques for continuous data and categorical data. As one can see, many techniques that are applicable to continuous data have no natural analogues in the categorical space.

When exploring the characteristics of a multi-dimensional continuous data set, we might begin by looking at one attribute at a time. We could compute the mean, percentiles, variance and skewness, or construct a box plot, histogram or nonparametric density

cap-shape	cap-surface	...	habitat	Class
convex	smooth		urban	poisonous
convex	smooth		grasses	edible
bell	smooth		meadows	edible
convex	scaly		urban	poisonous
convex	smooth		grasses	edible
...				

Table 1: Sample of the Mushroom Data Set from the UCI Machine Learning Repository [2].

function. This would give us an idea of the range and overall distribution of each attribute. However, with categorical data we can only look at the mode or an unordered histogram. With ordinal data (ordered categorical data), we may also be able to look at percentiles but for the most part the situation is similar to categorical data.

Other techniques that are extremely valuable in exploring continuous data including factor analysis techniques such as PCA, or multidimensional scaling can give us an idea about the variability of the data across all attributes. Multivariate techniques such as these are not even applicable in the categorical setting. Regardless of our final goal in analyzing a continuous data set, all of the above steps would help us understand its characteristics. On the other hand, when given a categorical data set many of these exploratory steps cannot naturally be extended to the new setting, leaving a huge “gap” as can be seen from Table 2. Thus, there is a need for elemental approaches for exploring the characteristics of a categorical data set.

In this paper, we propose a framework for categorical data analysis which seeks to address some of the limitations in analyzing categorical data sets. We seek to utilize underlying data characteristics for categorical data analysis, in the spirit of data-driven similarity measures. Specifically, we introduce the concept of *separability statistics*, which characterize the differences

	Continuous	Categorical
Single Attribute	Mean, Median, Box Plot, Histogram, Percentile, Variance, Skewness, Density Function	Mode, Histogram (no ordering)
Pairs of Attributes	Covariance, Scatter Plot, Correlation, 2-D Histogram, Density Function	Contingency Table, Correspondence Analysis, 2-D Histogram (no ordering)
Entire Space	PCA, Subspaces, MDS, LLE, SVD, ISOMAP, FastMAP	Subspaces, Data Cube
Other Techniques	Correlation Matrix ¹ , LDA	Correlation Matrix ¹ , Discriminant Correspondence Analysis

Table 2: Exploratory data analysis techniques for continuous and categorical data.

between a given instance and a labeled reference data set. Each statistic essentially represents a distance between an instance and the reference data set (i.e., the statistic allows mapping of the categorical data into a 1-dimensional continuous space). Therefore, using these statistics and a reference data set, one can map any collection of categorical instances (including those from the reference data set) to a multidimensional continuous space.

The key strength of the framework proposed in this paper is its ability to analyze a given data set with respect to a reference data set. In the transformed space, unseen instances similar to the reference data set will tend to occupy the same region that is occupied by instances from the reference data set. By contrast, instances that are different (we call this the *novel* class) will tend to be mapped to other regions, at least for some of the dimensions. This transformation of categorical data to continuous space can be utilized in practice for a variety of purposes.

Clustering and outlier detection require a similarity measure when applied to categorical data. In previous work [7], we have shown that the choice of similarity measure significantly affects overall performance. The proposed framework provides the capability to define a better similarity measure for a particular categorical data set; we will demonstrate this in the context of outlier detection, although one can extend this to other data mining tasks such as classification as well.

To illustrate the utility of separability statistics, let us consider a simple example. The *Mushroom* Data Set is a well-known categorical data set available from the UCI Machine Learning Repository [2]. This data set has 22 categorical attributes describing the various characteristics of a mushroom and a class which denotes whether a mushroom is edible or poisonous; the number of values taken by each of the attributes ranges between 2 and 12. While one can always explore the data set using techniques in Table 2 such

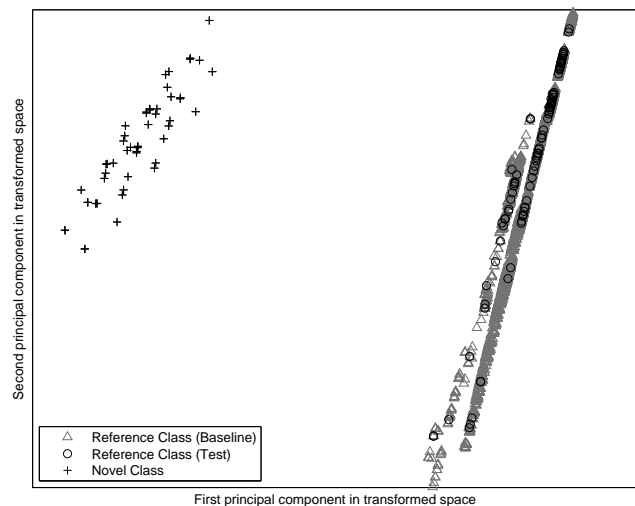


Figure 1: Visualization of the *Mushroom* data set using the proposed framework.

as an unordered histogram, these techniques are limited in what they can reveal about the *joint* distribution of the attributes. Table 1 shows the first few data instances in the Mushroom data set over a subset of the attributes. Using the methods to be discussed in this paper, this data set was mapped to a continuous space for visualization. Figure 1 shows the data instances in this transformed space with markers defined by the true labels: it is evident that the classes are well separated in this space. This allows the analyst to visually explore the classes in the Mushroom data set, which is not easy to do for the original categorical data set.

Key Contributions. The key contributions of this paper are as follows:

- We introduce the concept of separability statistics

¹A matrix which shows the intra- and inter-class correlation in a block structure [27, chap. 3].

in the context of exploratory categorical data analysis.

- We show how the statistics can be used as a way to map categorical data to continuous space given a labeled reference data set.
- This mapping enables visualizing of categorical data using techniques that are applicable to continuous data.
- We show that in the transformed space, the standard k -nn based outlier detection technique (designed for continuous space) works as well as the k -nn based outlier detection technique that uses the best of the similarity measures designed for categorical data.
- The proposed framework can be used to devise similarity measures best suited for a particular type of data set. We will demonstrate this in the context of outlier detection.

2 Related Work

Data with categorical attributes has been studied for a very long time, dating back at least a century when Karl Pearson [20, 21] introduced the χ^2 test for independence between categorical attributes. The traditional exploratory techniques used are contingency tables, the chi-square statistic, unordered histograms and pie charts [1]. Friendly [11] proposed several sophisticated statistical techniques such as *Sieve Diagrams* and *Mosaic Displays* to view k -way contingency tables, and *Multiple Correspondence Analysis* (MCA), to handle multivariate categorical data sets, though most techniques are limited to attributes that take few possible values. Fernandez [10] discusses several exploratory techniques for categorical data from a data mining perspective.

There have been a number of studies directed at categorical data in the visualization community [4, 6, 15, 16]. In particular, one direction in visualization has been to order the categories using the information present in the data [5, 19]. One such technique, called *Distance Quantification Classing* (DQC), was proposed by Rosario et al. [23] to order the categories present in a class variable in a categorical data set with respect to the predictor variables. None of the techniques directly address the problem of analyzing a categorical data set with respect to a reference data set, which is the focus of our paper.

A number of unsupervised learning algorithms have been proposed for categorical data, e.g. CLICKS by Zaki and Peters [30], CLOPE by Yang et al. [29], ROCK

by Guha et al. [14], CACTUS by Ganti et al. [12], COOLCAT by Barbará et al. [3], and other techniques by Gibson et al. [13] and Huang [17]. Most of these techniques use some notion of similarity when comparing instances. Similarity measures that are devised using the framework proposed in this paper can be plugged in to many such algorithms.

3 Separability Statistics

In this section we present a set of data-driven separability statistics that can be calculated for a given test data set with respect to a reference data set. Each statistic allows mapping of the categorical data into a 1-dimensional continuous space. The statistics are meant to differentiate instances in the reference data set from instances in other data sets. Since, for categorical data, the difference can be characterized in many different ways, a variety of separability statistics are possible. We only consider a few in this paper.

The discussion of the separability statistics is organized in the following manner: we will begin by discussing the intuition behind each of the statistics, including motivating examples, and then proceed to formally define the statistics. For now, let us refer to the four statistics as d_m , f_m , n_x and f_x . Let us also consider the simple categorical data set shown in Table 3, and the following two data instances: $\mathbf{y} = \langle a_1, b_1, c_{10}, d_1 \rangle$ and $\mathbf{z} = \langle a_3, b_2, c_{10}, d_5 \rangle$.

The statistic d_m essentially captures the extent to which a given instance has matching values with instances in the reference data set. This is driven by the intuition that an instance belonging to the same class as the reference class will, on average, have more matching values with the reference class than an instance belonging to a different class. The procedure to map categorical data to continuous space will be discussed in Section 4. For the purposes of the example being discussed with instances \mathbf{y} and \mathbf{z} , a brief outline is as follows: each statistic is computed by comparing a given instance with every instance in the reference data set, and then taking the average. For the instance \mathbf{y} , the values corresponding to the first and last rows of the data set would be 3 and 1 respectively. The final value of this statistic for the instances \mathbf{y} and \mathbf{z} is 1.5 and 0.6, respectively.

The statistic f_m takes into account the *frequency* of matching values between an instance and reference data set. One way to think of this statistic is as a frequency-weighted version of the statistic d_m . The key intuition here is that in addition to the importance of more matching values, instances belonging to the reference class will also tend to match on relatively frequent values, while instances not belonging to the

A	B	C	D
a_1	b_1	c_1	d_1
a_1	b_1	c_2	d_1
a_1	b_1	c_3	d_1
a_2	b_1	c_4	d_2
a_2	b_1	c_5	d_1
a_1	b_2	c_6	d_3
a_1	b_2	c_7	d_3
a_2	b_2	c_8	d_3
a_2	b_2	c_9	d_3
a_2	b_2	c_{10}	d_4

(a) Data set.

arity	A: 2	B: 2	C: 10	D: 4
frequency	$a_1: 5$	$b_1: 5$	$c_1: 1$	$d_1: 4$
	$a_2: 5$	$b_2: 5$	$c_2: 1$	$d_2: 1$
			$c_3: 1$	$d_3: 4$
			$c_4: 1$	$d_4: 1$
			$c_5: 1$	
			$c_6: 1$	
			$c_7: 1$	
			$c_8: 1$	
			$c_9: 1$	
			$c_{10}: 1$	

(b) Characteristics of attributes and values.

Table 3: A simple categorical data set with four attributes.

reference class will tend to match on infrequent values. This is important in situations where an attribute in the reference data set takes a very large number of values (e.g. the IP address in a network intrusion data set) thus making the odds of any match high. The value of this statistic for the instances \mathbf{y} and \mathbf{z} is 6.7 and 2.6, respectively.

The statistic n_x is a function of the *arity* of the mismatching attributes between an instance and a reference data set. In particular, the value of the statistic is higher when the mismatching attributes have lower arity, i.e. they take fewer values. The idea is that if an instance mismatches on an attribute that takes very few values across many instances in the reference class, then it is unlikely that it belongs to the class (simply because there are few opportunities to not match). The value of this statistic for the instances \mathbf{y} and \mathbf{z} is -5.45 and -7.90, respectively.

The statistic f_x looks at the frequency of mismatching attribute values between an instance and a reference data set. In a sense, this statistic is the “complement” of the f_m statistic and the intuition is also related to n_x ; if the frequency of mismatching values is high between a given instance and most members of the reference class is high, then this means the instance often mismatches with the reference class on values that are common in the reference class. Thus, it is unlikely that the instance belongs to the same class as the reference class. The value of this statistic for the instances \mathbf{y} and \mathbf{z} is -1.57 and -2.725, respectively.

The values assigned by the four statistics for instances \mathbf{y} and \mathbf{z} suggest that \mathbf{y} belongs to the reference class and \mathbf{z} does not. This is somewhat difficult to conclude just by looking at the instances and the reference data set, but by examining the underlying quantities behind the statistics one can see that it is indeed rea-

sonable to say that \mathbf{y} and \mathbf{z} belong to different classes. In particular, we have seen how the statistics map an instance between categorical space and continuous space based on several key underlying characteristics of the data set.

3.1 Formal Definition Table 4 lists the notation that will be used in the subsequent discussions.

T	Reference data set
D	Test data set
N	Size of reference data set
d	Number of attributes in T and D
a_i	i^{th} attribute ($1 \leq i \leq d$)
\mathcal{A}_i	Set of categorical values taken by a_i in T
n_i	Number of values taken by a_i ($= \mathcal{A}_i $)
$f_i(x)$	Number of times a_i takes value x in T

Table 4: Notation used in the paper.

Given a pair of categorical data instances $z \in D$ and $y \in T$, we define a partitioning of attribute set \mathcal{A} into \mathcal{A}_m and \mathcal{A}_x , such that, $z_i = y_i, \forall i \in \mathcal{A}_m$ and $z_i \neq y_i, \forall i \in \mathcal{A}_x$. \mathcal{A}_m denotes the set of matching attributes and \mathcal{A}_x denotes the set of mismatching attributes for the pair z, y .

We compute the following quantities for the pair z, y :

$$(3.1) \quad d_m = |\mathcal{A}_m|$$

$$(3.2) \quad f_m = \sum_{i \in \mathcal{A}_m} f_{z_i}$$

$$(3.3) \quad n_x = - \sum_{i \in \mathcal{A}_x} \frac{1}{n_i}$$

$$(3.4) \quad f_x = - \sum_{i \in \mathcal{A}_x} \left(\frac{1}{z_i} + \frac{1}{y_i} \right)$$

Thus, for every pair of categorical data instances $z \in D$ and $y \in T$ we have the following 4-tuple: $\langle d_m, f_m, n_x, f_x \rangle_{zy}$. For a test instance z we get a $|T| \times 4$ matrix of the above mentioned 4-tuple, denoted as:

$$(3.5) \quad \mathcal{M}_z = [\langle d_m, f_m, n_x, f_x \rangle_{zy}]_{\forall y \in T}$$

Let \vec{z}_k denote a row vector containing top k^{th} value for each column of \mathcal{M}_z , such that:

$$(3.6) \quad \vec{z}_k = \langle d_{mk}, f_{mk}, n_{xk}, f_{xk} \rangle$$

For a given value of k , we define a set of 4 statistics denoted as the row vector \vec{z}_k .

The reason to choose the top k^{th} value from each column of \mathcal{M}_z instead of a parameter independent value, such as the mean of the column, is to avoid issues due to multiple modes existing in the reference data set, T . If a very small value of k , such as 1, is chosen, the statistics can get affected by the presence of outliers in T . We have empirically observed that $5 \leq k \leq 15$ is a reasonable value of k for a variety of data sets. A set of statistics can be defined using multiple values of k to reduce the sensitivity on k .

Each of the four statistics mentioned in Equation 3.6 are motivated from the following observations in context of two instances $z_1, z_2 \in D$ and $y \in T$, such that z_1 is similar to instances (generated by the same distribution as T) in T while z_2 is different from the instances in T (not generated by the same distribution as T):

1. $d_{m|z_1y} > d_{m|z_2y}$.
2. $f_{m|z_1y} > f_{m|z_2y}$.
3. $f_{x|z_1y} > f_{x|z_2y}$.
4. $n_{x|z_1y} > n_{x|z_2y}$.

The above mentioned arguments indicate that if test instances $z \in D$ are transformed or mapped to \vec{z}_k , then the instances similar to T will map to the same region, while the instances different from T will map to a different region.

It should be noted that all of the above four observations might not necessarily hold true at the same time for a given data set. But one or more of them will likely hold true and hence by mapping the data into the joint space, one can distinguish between the two types of test instances.

Measure	$S_i(z_i, y_i)$	∞
<i>Overlap</i>	$= \begin{cases} 1 & \text{if } z_i = y_i \\ 0 & \text{otherwise} \end{cases}$	d_{mk}
<i>Goodall</i>	$= \begin{cases} \frac{f_i(z_i)(f_i(z_i)-1)}{N(N-1)} & \text{if } z_i = y_i \\ 0 & \text{otherwise} \end{cases}$	f_{mk}
<i>OF</i>	$= \begin{cases} 1 & \text{if } z_i = y_i \\ \frac{1}{1 + \log \frac{N}{f_i(z_i)} \times \log \frac{N}{f_i(y_i)}} & \text{otherwise} \end{cases}$	d_{mk}, f_{xk}
<i>Eskin</i>	$= \begin{cases} 1 & \text{if } z_i = y_i \\ \frac{n_i^2}{n_i^2 + 2} & \text{otherwise} \end{cases}$	d_{mk}, n_{xk}

Table 5: Similarity Measures for Categorical Attributes. Note that $S(z, y) = \sum_{i=1}^d S_i(z_i, y_i)$.

3.2 Relationship to Similarity Measures There have been several data driven similarity measures proposed for categorical data sets [7]. Table 5 lists four popular similarity measures that have been defined to measure similarity $S(z, y)$, between a pair of data instances.

We argue that the similarity of a test instance z to its k^{th} nearest neighbor in T using a data driven similarity measure, can be expressed as a function of one or more of the canonical statistics listed in Equation 3.6. Column 3 in Table 5 indicates the particular test statistic that corresponds to each similarity measure.

As an illustrative example, consider the similarity measure *Goodall* listed in Table 5. Let us consider a test instance z and the reference data set T . The *Goodall* similarity of z with an instance $y \in T$ can be written as:

$$\begin{aligned} S(z, y) &= \sum_{i \in \mathcal{A}_m} \frac{f_i(z_i)(f_i(z_i)-1)}{N(N-1)} + \sum_{i \in \mathcal{A}_x} 0 \\ &= \frac{1}{N(N-1)} \left(\sum_{i \in \mathcal{A}_m} f_i(z_i)^2 - \sum_{i \in \mathcal{A}_m} f_i(z_i) \right) \\ &\approx \frac{1}{N(N-1)} (f_m^2 - f_m) \quad (\text{See Eqn 3.2}) \end{aligned}$$

where \mathcal{A}_m and \mathcal{A}_x denote the set of attributes in which z and y match and mismatch, respectively. Note that we approximate $\sum_{i \in \mathcal{A}_m} f_i(z_i)^2$ with $(\sum_{i \in \mathcal{A}_m} f_i(z_i))^2$. The similarity of z to its k^{th} nearest neighbor in T , using the *Goodall*, is equal to the k^{th} largest value of $S(z, y) \forall y \in T$, and can be written as:

$$S^k(z, y) \approx \frac{1}{N(N-1)} (f_{mk}^2 - f_{mk})$$

Thus we have shown how the *Goodall* similarity measure

is related to the separability statistic f_{mk} . Similar relations can be shown for other similarity measures.

It may be argued that any similarity measure defined for categorical instances (such as the ones listed in Table 5 and others discussed in [7]) maybe used as a potential separability statistic in addition to the ones listed in Equation 3.6. But the statistics proposed in this paper are canonical and the similarity measures can be viewed as functions of one or more of the proposed statistics.

4 Mapping Data to Continuous Space

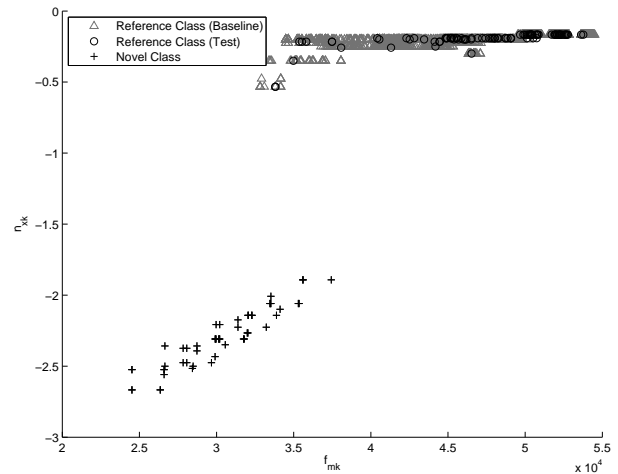
In this section we describe the process of mapping categorical data into a continuous space using the separability statistics discussed in Section 3.

For each categorical test instance in D , we first obtain the corresponding separability statistics as shown in Equation 3.6 with respect to the reference set T , using one or more values for k . The characteristic of this mapping is that test instances that belong to the reference class have lower values for each statistic than test instances that belong to the novel class. We denote the mapped test data set with \vec{D} .

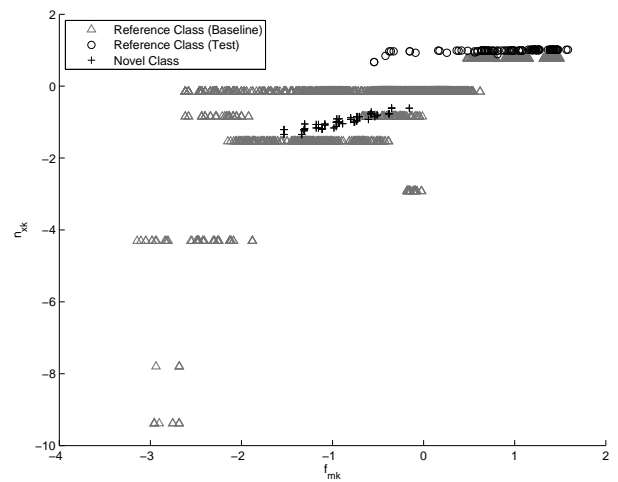
The reference data set T can also be mapped into a continuous space with respect to itself in the same manner as described above. We denote the mapped reference data set with \vec{T} . If the instances in T belong to a few dominant modes, one would expect the reference instances to map to similar values for each of the separability statistics.

Before further processing of the mapped data sets \vec{D} and \vec{T} , it is desirable to normalize the data, since the different statistics can take different ranges of values. Each column of the mapped data set \vec{D} is z -normalized to bring all statistics to the same scale. The mapped training data set \vec{T} is also normalized but in a slightly different manner; the difference being that the z -normalization of each column in \vec{T} is done using the column means and standard deviations obtained from the mapped test data set. This is important, because if the z -normalization of \vec{T} is done with respect to itself, the reference instances might have different normalized values for the statistics than the similar instances in the test set, which is not desirable.

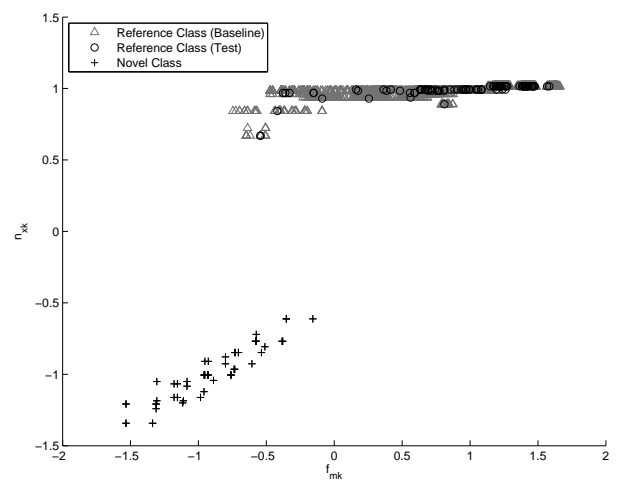
Figure 2 highlights the significance of the normalization, as described above, using the *Mushroom* data set. The plot 2(a) shows the mapped data using the raw statistics f_{mk} and n_{xk} . The range of f_{mk} statistic is $[2.0e+04, 5.5e+04]$ while the range of n_{xk} statistic is $[-0.53, -0.16]$. If the reference and test data sets are normalized independently, the reference instances are mapped to different values than the test instances belonging to the reference class as can be seen in Figure



(a) No normalization.



(b) Reference and test data sets independently normalized.



(c) Reference and test data sets normalized with respect to the test data set.

Figure 2: Plots of *Mushroom1* data set using statistics f_{mk} and f_{xk} .

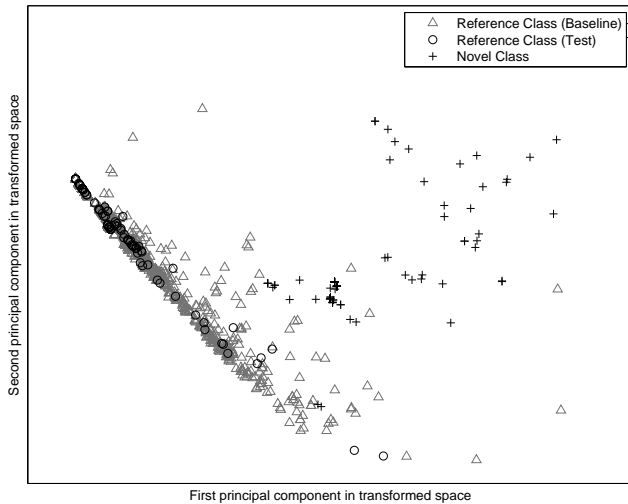


Figure 3: Visualization of KDD1 data set using a scatter plot.

2(b). If both reference and test data sets are normalized with respect to the test data set, the reference instances are normalized in the same fashion as the test instances belonging to the reference class, as can be seen in Figure 2(c) which is a scaled down version of the raw data in Figure 2(a).

5 Visualization

The separability statistics described in Section 3 allows for the visual exploration of any categorical data set. The data is first transformed as discussed in Section 4. Since the resulting data space is continuous, it is suitable for visualization. In particular, it allows the analyst to visually explore aspects such as separation between modes, size and the number of modes.

There are multiple ways to visualize the transformed continuous space, the simplest of which involve looking at pairs of dimensions or projections along specific subsets. Other mechanisms can be used to visualize continuous space such as tours in the GGobi system [26] and those in the Orca system [25]. We refer the reader to the recent work by Wickham et al. [28] and Lawrence et al. [18] for a discussion of high-dimensional data visualization systems. In this paper, we will discuss two ways, one which utilizes dimensionality reduction and another with histograms.

5.1 Two-dimensional Scatter Plots. In order to reduce the number of dimensions to two for the purpose of visualization, we will use the well-known principal components analysis (PCA) technique [8]. The role of PCA here is to give more emphasis to the statistics

that exhibit more separability and filter out those that are close to unimodal. For this paper we have chosen PCA for its simplicity, however, there are other dimensionality reduction techniques that may be better suited for this task; we will not discuss other techniques since they are out of the scope of this paper.

To illustrate the use of our proposed framework for visualization, we will consider an example with a real data set. The data set has been partitioned into a labeled reference data set and a test data set. The test data set is then mapped to continuous space using the procedure discussed in Section 4. The resulting space \bar{T} is 4-dimensional and each column is normalized; PCA is applied to the data set and the leading two principal components are preserved. The data set \bar{T} is then projected on to the two leading principal components resulting in a 2-dimensional data set, with the number of rows being the number of test instances.

We can now visually explore the two-dimensional space; in this case we will use a scatter plot. The idea is that instances that have similar values for the statistics will end up in the same region of the plot, while those that have different values will be in different regions. The key observation is that the instances that have different values are likely to be from a different class than the reference data set.

Figure 3 shows the scatter plot for the KDD1 data set, which has 29 attributes, some of which take hundreds of values. Note that the labels of the test data set were examined only *after* the data was mapped to this space. The test data set contained instances from the reference class as well as instances that did not belong to the reference class. It is evident that the separability statistics were effective in distinguishing the classes. In particular, note that the instances belonging to the reference class were mapped to the same region, even though some of these instances came from the test class for which the label was unknown during the analysis. Another advantage of a dimensionality reduction technique is that it returns a linear combination of the statistics which is optimal in some sense. The weights from the linear combination can then be used to design a similarity measure for the data set (this aspect will be further discussed in Section 6.1).

5.2 Histograms. Since the separability statistics are directly capturing important characteristics of the underlying data, it is very useful to examine their distribution using a histogram. In this section, we will discuss exploring the distribution of a single statistic. As stated earlier, if a statistic assigns different values to a set of instances compared to the reference class, they are likely to be from a different class than the reference data set.

Therefore, the distribution of the statistic will be unimodal (with low variance) when all instances are from the same class. One way to examine the distribution of a statistic is using a histogram. The histogram will essentially show to what extent the distribution departs from a low variance unimodal behavior.

Figure 4 shows histograms of the four statistics for the KDD1 data set (the labels were examined only after the histogram was constructed). In this case, it is apparent that the distribution of all the statistics are multi-modal with high variance. If we were to generate this plot without knowing the labels, we would observe that the f_m and d_m statistics exhibited a high degree of multi-modality, while the other two statistics were somewhat multi-modal. Therefore, f_m and d_m would be considered the best separating statistics for this data set. Looking at Figure 4, taking the labels into account we see that this is indeed the case. Based on these histograms, the conclusion for the KDD1 data set would be that the reference class can be distinguished from other classes using properties related to the d_m statistic (more matching values) and the f_m statistic (more matches on frequent values).

6 Utility of Separability Statistics for Outlier Detection

In this section we illustrate the utility of the separability statistics in semi-supervised outlier detection. Here the objective is to separate outliers from normal instances in a given test data set, with respect to a reference (training) data set which is assumed to contain only normal instances.

We use a nearest neighbor based outlier detection technique (kNN) [22, 27] which assigns the outlier score of a test instance as equal to the distance of the test instance to its k^{th} nearest neighbor in the reference data set. The distance can be computed using any distance measure. If a measure computes similarity, the outlier score of a test instance is inverse of the similarity to its k^{th} nearest neighbor.

We experimented with two kNN based outlier detection techniques using the separability statistics. In the first variation (denoted as $kNNEuc$), we assign an outlier score to each test instance in \vec{D} using \vec{T} as the reference data, using *Euclidean* distance as the distance measure.

In the second variation of kNN (denoted as $kNNPCA$), we use *Principal Component Analysis* (PCA) to project the mapped data sets, \vec{D} and \vec{T} , to a lower dimensional space. PCA is performed on the mapped test data set \vec{D} . The top principal components that capture 90% of the variance in the test data are chosen. Both test and reference data sets are projected

along these top principal components. Outlier scores are assigned to test instances using *Euclidean* distance in this projected space. Both variations combine the four statistics when computing distance between instances.

The motivation behind using PCA is that the statistics that can discriminate between normal and outliers in the test data tend to have higher variance than the statistics that do not discriminate between normal and outliers. By using PCA, we can capture the statistics with greater discriminative power.

To evaluate the performance of any technique, we count the number of true outliers in the top n portion of the sorted outliers scores of the test instances, where n is the number of actual outliers. Let o be the number of actual outliers in the top p predicted outliers. The accuracy of the algorithm is measured as $\frac{o}{n}$.

We compare the two variants described above with 14 different categorical similarity measures on several publicly available data sets. Four of these similarity measures are listed in Table 5. The other ten measures have been developed in different contexts, and have been evaluated in [7]. The details of the data sets are summarized in Table 6. Fourteen of these data sets are based on the data sets available at the UCI Machine Learning Repository [2], while two are based on network data generated by SKAION Corporation for the ARDA information assurance program [24]. Nine of these data sets were purely categorical while seven (kd1,kd2,kd3,kd4,sk1,sk2,cen) had a mix of continuous and categorical attributes. Continuous variables were discretized using the MDL method [9]. Another possible way to handle a mixture of attributes is to compute the similarity for continuous and categorical attributes separately, and then do a weighted aggregation. In this study we converted the continuous attributes to categorical to simplify comparative evaluation.

For each test data set there is a corresponding normal reference data set, and a labeled test data set. The results are summarized in Table 7. The row *stt* denotes the performance of kNN when using the best separability statistic as the only attribute. The best statistic is indicated in the last row. We make several observations from the results in Table 7.

The performance of the similarity measures depends on the data set, which is expected, since the measures are data-driven. This also indicates that the ability of the underlying statistic to distinguish between normal and outliers depends on the data set. Since each similarity measure is a function of one statistic, we observe that the similarity measure which uses the best statistic for a given data set, is generally the best performer.

The performance of $kNNEuc$ technique (using all

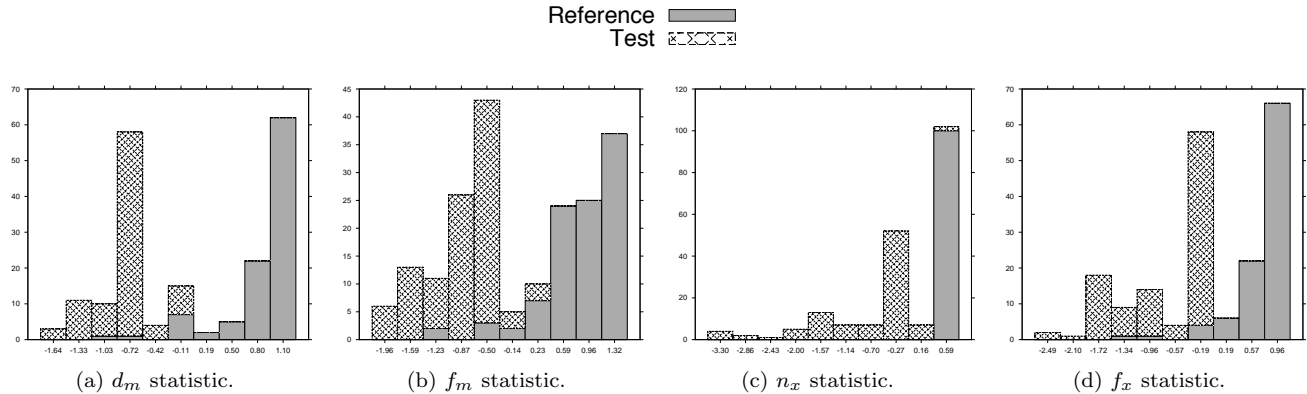


Figure 4: Visualization of KDD1 data set using histograms.

	cr1	cr2	cn1	cn2	kd1	kd2	kd3	kd4	sk1	sk2	ms1	ms2	cen	bal	ttt	aud
d	6	6	42	42	29	29	29	29	10	10	21	21	10	4	9	16
$ T $	904	944	3055	3055	1000	1000	6007	6007	2182	1429	3208	2916	2120	106	316	73
$ D $	759	715	1100	550	1100	1100	1100	1100	1298	1177	1100	1100	2321	308	341	77

Table 6: Description of public data sets used for experimental evaluation. Each test data sets contains normal and outliers in ratio 10:1.

separability statistics) is one of the best on average. This result shows that when all statistics are used together, the performance can often be better than using them individually, though in several cases the performance deteriorates considerably when all statistics are used (such as for $cn2$ and $sk2$).

The $kNNPCA$ technique performs better on average than all 14 data driven similarity measures and the $kNNEuc$ technique. This shows that PCA is able to capture a better combination of the separability statistics automatically than captured by the similarity measures. Moreover, it also shows that using all statistics may not be optimal for several data sets, and an optimal subset is required to be selected. For some data sets we observe that $kNNPCA$ does not perform as well as using a single best discriminating statistic which is shown in the row stt (the corresponding statistic is shown in row ind). This shows that PCA might not always be able to determine the best combination of the statistics.

The performance of the best statistic (row stt) is the best for most of the data sets. In some cases, such as $sk1$ and $ms2$, the combination of statistics (using *Euclidean* distance or PCA) outperforms the single best statistic.

6.1 Designing a Better Similarity Measure The results in Table 7 show that for many data sets, a combination of the separability statistics can result in better performance than using them individually. PCA is one way to obtain such a combination, but as the

results indicate, it might not always provide the optimal combination. If a labeled validation data set is present, one can visually inspect the histograms for different statistics, and select the ones that provide maximum separation between the normal and outliers. We argue that using this approach we can arrive at an optimal subset of separability statistics. A similarity measure can then be designed to use this subset.

To verify the above hypothesis we conducted the following experiment. We selected data sets $sk1$ and $sk2$ from Table 7. For each data set, the test data is split into equal sized validation and test sets. We first map the validation set into continuous space and analyze the histograms for each separability statistic, making use of the labels for the validation instances. We then select a subset of the statistics that best separate the normal points and outliers. Figures 5 and 6 show the per-statistic histograms for $sk1$ and $sk2$ data sets, respectively.

We observe that for $sk1$, statistics 1 and 3 (d_{mk} and f_{xk}) show maximum separability between normal and outliers in the corresponding validation data set. Similarly, for $sk2$, statistics 3 and 4 (f_{xk} and n_{xk}) show maximum separability between normal and outliers in the validation data set. We then apply the *Euclidean* distance based kNN technique on the test data set using the best subset of statistics.

Table 8 summarizes the performance of kNN using different similarity measures and the performance of kNN using the best subset of statistics on the two data

	cr1	cr2	cn1	cn2	kd1	kd2	kd3	kd4	sk1	sk2	ms1	ms2	cen	bal	ttt	aud	<i>Avg</i>
ovr	0.16	0.06	0.38	0.14	0.88	0.97	0.90	0.90	0.68	0.44	1.00	0.96	0.11	0.04	0.23	0.43	0.52
gd4	0.45	0.65	0.10	0.06	0.79	0.93	0.90	0.90	0.12	0.08	0.78	0.93	0.07	0.07	0.52	0.29	0.48
of	0.54	0.58	0.64	0.16	0.82	0.94	0.85	0.78	0.68	0.42	1.00	0.96	0.19	0.04	0.29	0.43	0.58
esk	0.51	0.54	0.39	0.14	0.88	0.96	0.90	0.90	0.68	0.30	1.00	0.96	0.23	0.04	0.23	0.43	0.57
iof	0.14	0.46	0.51	0.16	0.70	0.87	0.73	0.81	0.25	0.17	1.00	0.95	0.09	0.07	0.87	0.29	0.51
lin	0.00	0.00	0.29	0.26	0.86	0.96	0.90	0.88	0.75	0.60	1.00	0.97	0.09	0.21	0.45	0.29	0.53
lin1	0.42	0.65	0.28	0.24	0.91	0.95	0.82	0.09	0.72	0.39	1.00	0.97	0.18	0.00	0.23	0.29	0.51
gd1	0.00	0.00	0.20	0.22	0.81	0.90	0.00	0.01	0.69	0.30	1.00	0.81	0.12	0.25	0.35	0.43	0.38
gd2	0.54	0.71	0.62	0.22	0.78	0.89	0.18	0.11	0.69	0.55	1.00	0.96	0.16	0.04	0.32	0.43	0.51
gd3	0.01	0.00	0.24	0.18	0.81	0.91	0.00	0.11	0.69	0.41	1.00	0.96	0.16	0.14	0.32	0.43	0.40
smv	0.00	0.00	0.07	0.16	0.00	0.00	0.00	0.00	0.34	0.07	0.00	0.00	0.07	0.21	0.35	0.00	0.08
gmb	0.57	0.68	0.67	0.24	0.72	0.91	0.79	0.85	0.20	0.20	1.00	0.90	0.15	0.04	0.35	0.43	0.54
brb	0.12	0.52	0.43	0.14	0.91	0.96	0.90	0.90	0.66	0.36	1.00	0.96	0.10	0.18	0.87	0.29	0.58
anb	0.00	0.02	0.15	0.14	0.58	0.78	0.69	0.22	0.51	0.09	1.00	0.88	0.21	0.14	0.39	0.29	0.38
euc	0.55	0.65	0.18	0.14	0.89	0.96	0.90	0.90	0.66	0.26	1.00	0.96	0.18	0.11	0.35	0.71	0.59
pca	0.55	0.72	0.18	0.14	0.90	0.96	0.90	0.90	0.71	0.42	1.00	0.95	0.18	0.11	0.39	0.71	0.61
stt	0.54	0.65	0.38	0.16	0.91	0.98	0.90	0.90	0.71	0.73	1.00	0.96	0.18	0.14	0.45	0.71	0.64
	f_{xk}	f_{mk}	d_{mk}	f_{mk}	f_{xk}	f_{xk}	d_{mk}	d_{mk}	f_{xk}	f_{xk}	d_{mk}	d_{mk}	f_{xk}	f_{mk}	f_{mk}	d_{mk}	
<i>Avg</i>	0.30	0.40	0.34	0.17	0.77	0.87	0.66	0.60	0.57	0.34	0.93	0.88	0.15	0.11	0.41	0.40	

Table 7: Performance of similarity measures and separability statistics on public data sets using kNN ($k = 10$).

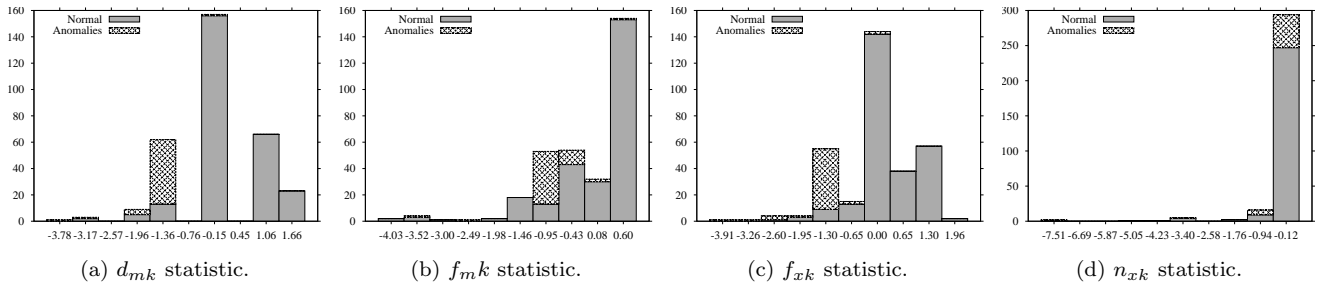


Figure 5: Histograms of separability statistics for data set $sk1$.

	$sk1$		$sk2$	
	Val.	Test	Val.	Test
<i>ovr</i>	0.71	0.69	0.71	0.69
<i>gd4</i>	0.16	0.14	0.12	0.16
<i>of</i>	0.78	0.74	0.82	0.74
<i>esk</i>	0.68	0.72	0.73	0.70
d_{mk}	0.71	0.69	0.71	0.69
f_{mk}	0.16	0.18	0.43	0.41
f_{xk}	0.75	0.78	0.79	0.69
n_{xk}	0.56	0.53	0.79	0.49
<i>euc</i>	0.73	0.61	0.84	0.78
<i>pca</i>	0.75	0.63	0.84	0.78
eucs	0.84	0.82	0.84	0.82

Table 8: Outlier detection performance for $sk1$ and $sk2$ data sets ($k = 10$). Row *eucs* shows the results using the best subset of statistics.

sets. The results show that while none of the statistics individually perform as well (maximum accuracy is 0.78 for f_{xk} in $sk1$ and 0.69 for f_{xk} in $sk2$), the combination of the two best statistics (from the histograms as well as results of statistics on the validation data set), has accuracy of 0.82 for both data sets.

The results also indicate that the other two combination methods, *viz.*, *Euclidean* and *PCA*, are slightly worse than the optimal combination, but still outperform all similarity measures as well as the individual statistics.

Thus, given a validation data set, a better subset of statistics can be selected by either using the histograms or by observing the results of individual statistics on the validation data set.

7 Concluding Remarks and Future Research Directions

This paper presents a framework to analyze categorical data. It is clear from the discussion in the previous sections that there is a tremendous gap between exploratory data analysis techniques for continuous and categorical data sets. This paper is an attempt towards

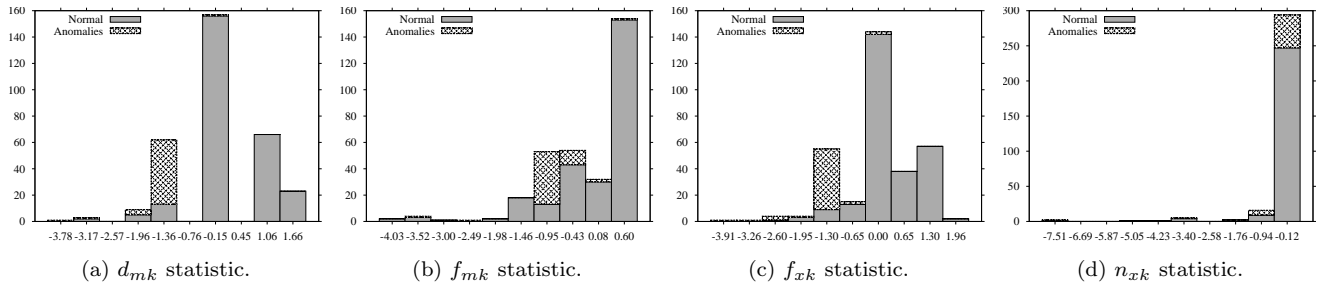


Figure 6: Histograms of separability statistics for data set *sk2*.

bridging this gap. By mapping categorical data to continuous space, we open up the possibility of utilizing exploratory techniques that are available for continuous data to be applied to categorical data. The key strength of the proposed framework is its ability to analyze a given test data set with respect to a reference data set. We have demonstrated how this property can be used for visualization as well as outlier detection. In both applications, the framework is used to distinguish between instances belonging to the reference class(es) against the instances belonging to a novel class. Visualization allows an analyst to understand the data, set optimal parameters (such as number of nearest neighbors k), as well as choose or design optimal similarity measures using the proposed statistics. We believe that this framework can be extended in several directions, and discuss some future directions for research here.

The separability statistics, discussed in Section 3, are inspired from different similarity measures that have been proposed for categorical data. Many other such canonical statistics can be developed, which can be used to distinguish between instances that belong to the reference class against the other instances, e.g., a statistic that captures the correlation between different attributes.

Note that each of the separability statistics as well as their combinations can serve as distance/similarity measures. We showed that one can select an appropriate subset of separability statistics (or their linear combination, e.g. using PCA) in a supervised setting. This opens up the possibility for devising entirely new distance/similarity measures for categorical data sets.

In this paper we have used two standard visualization techniques, viz., histograms and 2-D plots of data projected on top two principal components. Other visualization and exploratory techniques that are applied to continuous data (see Table 2, [18], [25]), can also be applied to the mapped data.

We have demonstrated the discriminative power of the framework in the context of outlier detection, but

one can extend it for other data mining tasks such as classification and clustering. Moreover, the concept of analyzing a test data set with respect to a reference data set can also be extended to continuous data. Specifically, using a set of separability statistics (similar to the ones proposed in Section 3), continuous data can also be analyzed in the same framework.

Another possible extension to the proposed framework is to analyze a categorical data set with respect to itself. If the data mostly contains instances belonging to one or a few dominant modes, and a few outliers, the outliers should, in principle, appear different than the normal instances in the mapped space. Thus, the framework can be used for tasks such as unsupervised outlier detection or noise removal.

References

- [1] A. Agresti. *Categorical Data Analysis*. John Wiley & Sons, 2003.
- [2] A. Asuncion and D. J. Newman. UCI machine learning repository. [<http://archive.ics.uci.edu/ml>]. Irvine, CA: University of California, 2007.
- [3] D. Barbará, Y. Li, and J. Couto. COOLCAT: an entropy-based algorithm for categorical clustering. In *CIKM '02: Proceedings of the eleventh international conference on Information and knowledge management*, pages 582–589, New York, NY, USA, 2002. ACM.
- [4] F. Bendix, R. Kosara, and H. Hauser. Parallel sets: Visual analysis of categorical data. In *IEEE Symposium on Information Visualization*, page 18, Los Alamitos, CA, USA, 2005. IEEE Computer Society.
- [5] A. Beygelzimer, C.-S. Perng, and S. Ma. Fast ordering of large categorical datasets for better visualization. In *Proceedings of the seventh ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 239–244. ACM, 2001.

- [6] J. Blasius and M. Greenacre. *Visualization of Categorical Data*. Academic Press, 1998.
- [7] S. Boriah, V. Chandola, and V. Kumar. Similarity measures for categorical data: A comparative evaluation. In *SDM 2008: Proceedings of the eighth SIAM International Conference on Data Mining*, pages 243–254, 2008.
- [8] J. W. Demmel. *Applied Numerical Linear Algebra*. Society for Industrial and Applied Mathematics, Philadelphia, PA, USA, 1997.
- [9] U. M. Fayyad and K. B. Irani. Multi-interval discretization of continuous-valued attributes for classification learning. In *Proceedings of the 13th International Joint Conference on Artificial Intelligence*, pages 1022–1029, San Francisco, CA, 1993. Morgan Kaufmann.
- [10] G. Fernandez. *Data mining using SAS applications*. Chapman & Hall/CRC, Boca Raton, FL, USA, 2003.
- [11] M. Friendly. *Visualizing Categorical Data*. SAS Publishing, 2000.
- [12] V. Ganti, J. Gehrke, and R. Ramakrishnan. CACTUS—clustering categorical data using summaries. In *KDD '99: Proceedings of the fifth ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 73–83, New York, NY, USA, 1999. ACM Press.
- [13] D. Gibson, J. Kleinberg, and P. Raghavan. Clustering categorical data: an approach based on dynamical systems. *The VLDB Journal*, 8(3):222–236, 2000.
- [14] S. Guha, R. Rastogi, and K. Shim. ROCK: A robust clustering algorithm for categorical attributes. *Information Systems*, 25(5):345–366, 2000.
- [15] P. Hoffman and G. Grinstein. A survey of visualizations for high-dimensional data mining. In U. M. Fayyad, G. G. Grinstein, and A. Wierse, editors, *Information Visualization in Data Mining and Knowledge Discovery*, chapter 2, pages 47–82. Morgan Kaufmann, 2002.
- [16] H. Hofmann. Exploring categorical data: interactive mosaic plots. *Metrika*, 51(1):11–26, 2000.
- [17] Z. Huang. Extensions to the k-means algorithm for clustering large data sets with categorical values. *Data Mining and Knowledge Discovery*, 2(3):283–304, 1998.
- [18] M. Lawrence, H. Wickham, D. Cook, H. Hofmann, and D. Swayne. Extending the GGobi pipeline from R. *Computational Statistics*, 2008.
- [19] S. Ma and J. L. Hellerstein. Ordering categorical data to improve visualization. In *IEEE Information Visualization Symposium Late Breaking Hot Topics*, pages 15–18. IEEE, 1999.
- [20] K. Pearson. *On the Theory of Contingency and Its Relation to Association and Normal Correlation*. Dulau and Co., 1904.
- [21] K. Pearson. On the general theory of multiple contingency with special reference to partial contingency. *Biometrika*, 11(3):145–158, 1916.
- [22] S. Ramaswamy, R. Rastogi, and K. Shim. Efficient algorithms for mining outliers from large data sets. In *Proceedings of the ACM SIGMOD International Conference on Management of Data*, pages 427–438. ACM Press, 2000.
- [23] G. Rosario, E. Rundensteiner, D. Brown, M. Ward, and S. Huang. Mapping nominal values to numbers for effective visualization. *Information Visualization*, 3(2):80–95, 2004.
- [24] SKAION Corporation. SKAION intrusion detection system evaluation data. [<http://www.skaion.com/news/rel20031001.html>].
- [25] P. Sutherland, A. Rossini, T. Lumley, N. Lewin-Koh, J. Dickerson, Z. Cox, and D. Cook. Orca: A visualization toolkit for high-dimensional data. *Journal of Computational and Graphical Statistics*, 9(3):509–529, 2000.
- [26] D. F. Swayne, D. T. Lang, A. Buja, and D. Cook. GGobi: evolving from XGobi into an extensible framework for interactive data visualization. *Computational Statistics & Data Analysis*, 43(4):423–444, 2003.
- [27] P.-N. Tan, M. Steinbach, and V. Kumar. *Introduction to Data Mining*. Addison-Wesley, Boston, MA, 2006.
- [28] H. Wickham, M. Lawrence, D. Cook, A. Buja, H. Hofmann, and D. Swayne. The plumbing of interactive graphics. *Computational Statistics*, 2008.
- [29] Y. Yang, X. Guan, and J. You. CLOPE: a fast and effective clustering algorithm for transactional data. In *KDD '02: Proceedings of the eighth ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 682–687, New York, NY, USA, 2002. ACM.
- [30] M. J. Zaki and M. Peters. CLICKS: Mining subspace clusters in categorical data via k-partite maximal cliques. In *ICDE 2005: Proceedings of the 21st International Conference on Data Engineering*, pages 355–356, 2005.