Similarity Measures for Categorical Data: A Comparative Evaluation

Shyam Boriah Varun Chandola Vipin Kumar

Department of Computer Science and Engineering
University of Minnesota

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

Abstract

Measuring similarity or distance between two entities is a key step for several data mining and knowledge discovery tasks. The notion of similarity for continuous data is relatively well-understood, but for categorical data, the similarity computation is not straightforward. Several data-driven similarity measures have been proposed in the literature to compute the similarity between two categorical data instances but their relative performance has not been evaluated. In this paper we study the performance of a variety of similarity measures in the context of a specific data mining task: outlier detection. Results on a variety of data sets show that while no one measure dominates others for all types of problems, some measures are able to have consistently high performance.

1 Introduction

Measuring similarity or distance between two data points is a core requirement for several data mining and knowledge discovery tasks that involve distance computation. Examples include clustering (k-means), distance-based outlier detection, classification (knn, SVM), and several other data mining tasks. These algorithms typically treat the similarity computation as an orthogonal step and can make use of any measure.

For continuous data sets, the *Minkowski Distance* is a general method used to compute distance between two multivariate points. In particular, the *Minkowski Distance* of order 1 (*Manhattan*) and order 2 (*Euclidean*) are the two most widely used distance measures for continuous data. The key observation about the above measures is that they are independent of the underlying data set to which the two points belong. Several data-driven measures, such as *Mahalanobis Distance*, have also been explored for continuous data.

The notion of similarity or distance for categorical data is not as straightforward as for continuous data. The key characteristic of categorical data is that the different values that a categorical attribute takes are not inherently ordered. Thus, it is not possible to directly compare two different categorical values. The simplest way to find similarity between two categorical attributes is to assign a similarity of 1 if the values are identical and a similarity of 0 if the values are not identical. For two multivariate categorical data points, the similarity between them will be directly proportional to the number of attributes in which they match. This simple measure is also known as the *overlap* measure in the literature [33].

One obvious drawback of the *overlap* measure is that it does not distinguish between the different values taken by an attribute. All matches, as well as mismatches, are treated as equal. For example, consider a categorical data set D, defined over two attributes: *color* and *shape*. Let *color* take 3 possible values in D: {*red, blue, green*} and *shape* take 3 possible values in D: {*square, circle, triangle*}. Table 1 summarizes the frequency of occurrence for each possible combination in D.

		shape			
		square	*circle*	*triangle*	Total
color	*red*	30	2	3	35
	blue	25	25	0	50
	green	2	1	2	5
	Total	57	28	5	

Table 1: Frequency Distribution of a Simple 2-D Categorical Data Set

The *overlap* similarity between two instances (*green,square*) and (*green,circle*) is $\frac{1}{3}$. The *overlap* similarity between (*blue,square*) and (*blue,circle*) is also $\frac{1}{3}$. But the frequency distribution in Table 1 shows that while (*blue,square*) and (*blue,circle*) are frequent combinations, (*green,square*) and (*green,circle*) are very rare

combinations in the data set. Thus, it would appear that the *overlap* measure is too simplistic in giving equal importance to all matches and mismatches. Although there is no inherent ordering in categorical data, the previous example shows that there is other information in categorical data sets that can be used to define what should be considered more similar and what should be considered less similar.

This observation has motivated researchers to come up with data-driven similarity measures for categorical attributes. Such measures take into account the frequency distribution of different attribute values in a given data set to define similarity between two categorical attribute values. In this paper, we study a variety of similarity measures proposed in diverse research fields ranging from statistics to ecology as well as many of their variations. Each measure uses the information present in the data uniquely to define similarity.

Since we are evaluating data-driven similarity measures it is obvious that their performance is highly related to the data set that is being analyzed. To understand this relationship, we first identify the key characteristics of a categorical data set. For each of the different similarity measure that we study, we analyze how it relates to the different characteristics of the data set.

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

- We bring together fourteen different categorical measures from different fields and study them together in a single context. Many of these measures have not been investigated outside the domain they were introduced in, and not compared with other measures.

- We classify the categorical measures in three different ways based on how they utilize information in the data.

- We evaluate the various similarity measures for categorical data on a wide variety of benchmark data sets. In particular, we show the utility of data-driven measures for the problem of determining similarity with categorical data.

- We also propose a number of new measures that are either variants of other previously proposed measures, or derived from previously proposed similarity frameworks. The performance of some of the measures we propose is among the best performance of all the measures we study.

1.2 Organization of the Paper. The rest of the paper is organized as follows. We first mention all related efforts in the study of similarity measures in Section 2. In section 3, we identify various characteristics of categorical data that are relevant to this study. We then introduce the 14 different similarity measures that are studied in this paper in Section 4. We describe our experimental setup, evaluation methodology and the results on public data sets in Section 5.

2 Related Work

Sneath and Sokal discuss categorical similarity measures in some detail in their book [32] on numerical taxonomy. They were among the first to put together and discuss many of the measures discussed in their book. At the time, two major concerns were (1) biological relevance, since numerical taxonomy was mainly concerned with taxonomies from biology, ecology, etc., and (2) computation efficiency since computational resources were limited and scarce. Nevertheless, many of the observations made by Sneath and Sokal are quite relevant today and offer key insights into many of the measures.

There are several books [2, 18, 16, 21] on cluster analysis that discuss the problem of determining similarity between categorical attributes. However, most of these books do not offer solutions to the problem or discuss the measures in this paper, and the usual recommendation is to binarize the data and then use binary similarity measures.

Wilson and Martinez [36] performed a detailed study of heterogeneous distance functions (for data with categorical and continuous attributes) for instance-based learning. The measures in their study are based upon a supervised approach where each data instance has class information in addition to a set of categorical/continuous attributes. Measures discussed in this paper are orthogonal to the ones proposed in [36] since supervised measures determine similarity based on class information, while data-driven measures determine similarity based on the data distribution. In principle, both ideas can be combined.

There have been a number of new data mining techniques for categorical data that have been proposed recently. Some of them use notions of similarity which are neighborhood-based [15, 4, 8, 26, 1, 22], or incorporate the similarity computation into the learning algorithm [13, 17, 12]. Neighborhood-based approaches use some notion of similarity (usually the *overlap* measure) to define the neighborhood of a data instance, while the measures we study in this paper are directly used to determine similarity between a pair of data instances; hence, we see the measures discussed in this paper as being useful to compute the neighborhood of a point and

neighborhood-based measures as meta-similarity measures. Since techniques which embed similarity measures into the learning algorithm do not explicitly define general categorical similarity measures, we do not discuss them in this paper.

Jones and Furnas [20] studied several similarity measures in the field of information retrieval. In particular, they performed a geometric analysis on continuous measures in order to reveal important differences which would affect retrieval performance. Noreault et al. [25] also studied measures in information retrieval with the goal of generalizing effectiveness based on empirically evaluating the performance of the measures. Another comparative empirical evaluation–for determining similarity between fuzzy sets–was performed by Zwick et al. [37], followed by several others [27, 35].

3 Categorical Data

Categorical data (also known as nominal or qualitative multi-state data) has been studied for a long time in various contexts. As mentioned earlier, computing similarity between categorical data instances is not straightforward owing to the fact that there is no explicit notion of ordering between categorical values. To overcome this problem, several data-driven similarity measures have been proposed for categorical data. The behavior of such measures directly depends on the data. In this section, we identify the key characteristics of a categorical data set that can potentially affect the behavior of a data-driven similarity measure.

For the sake of notation, consider a categorical data set D containing N objects, defined over a set of d categorical attributes where A_k denotes the k^{th} attribute. Let the attribute A_k take n_k values in the given data set that are denoted by the set $\mathcal{A}_k$. We also use the following notation:

- $f_k(x)$: The number of times attribute A_k takes the value x in the data set D. Note that if $x \notin \mathcal{A}_k$, $f_k(x) = 0$
- $\hat{p}_k(x)$: The sample probability of attribute A_k to take the value x in the data set D. The sample probability is given by

$$\hat{p}_k(x) = \frac{f_k(x)}{N}$$

- $p_k^2(x)$: Another probability estimate of attribute A_k to take the value x in a given data set, given by

$$p_k^2(x) = \frac{f_k(x)(f_k(x)-1)}{N(N-1)}$$

3.1 Characteristics of Categorical Data.

Since this paper discusses data-driven similarity measures for categorical data, a key task is to identify the characteristics of a categorical data set that affect the behavior of such a similarity measure. We enumerate the characteristics of a categorical data set below:

- *Size of Data*, N. As we will see later, most measures are typically invariant of the size of the data, though there are some measures (e.g. *Smirnov*) that do make use of this information.
- *Number of attributes*, d. Most measures are invariant of this characteristic, since they typically normalize the similarity over the number of attributes. But in our experimental results we observe that the number of attributes does affect the performance of the outlier detection algorithms.
- *Number of values taken by each attribute*, n_k. A data set might contain attributes that take several values and attributes that take very few values. For example, one attribute might take several hundred possible values, while another attribute might take very few values. A similarity measure might give more importance to the second attribute, while ignoring the first one. In fact, one of the measures discussed in this paper (*Eskin*) behaves exactly like this.
- *Distribution of* $f_k(x)$. This refers to the distribution of frequency of values taken by an attribute in the given data set. In certain data sets an attribute might be distributed uniformly over the set $\mathcal{A}_k$, while in others the distribution might be skewed. A similarity measure might give more importance to attribute values that occur rarely, while another similarity measure might give more importance to frequently occurring attribute values.

4 Similarity Measures for Categorical Data

The study of similarity between data objects with categorical variables has had a long history. Pearson proposed a *chi-square statistic* in the late 1800s which is often used to test independence between categorical variables in a contingency table. Pearson's chi-square statistic was later modified and extended, leading to several other measures [28, 24, 7]. More recently, however, the *overlap* measure has become the most commonly used similarity measure for categorical data. Its popularity is perhaps related to its simplicity and easy of use. In this section, we will discuss the *overlap* measure and several data-driven similarity measures for categorical data.

Note that we have converted measures that were originally proposed as distance to similarity measures in order to make the measures comparable in this study. The measures discussed henceforth will all be in the context of similarity, with distance measures being converted using the formula:

$$sim = \frac{1}{1 + dist}$$

Almost all similarity measures assign a similarity value between two data instances X and Y belonging to the data set D (introduced in Section 3) as follows:

$$S(X, Y) = \sum_{k=1}^{d} w_k S_k(X_k, Y_k) \tag{4.1}$$

where $S_k(X_k, Y_k)$ is the per-attribute similarity between two values for the categorical attribute A_k. Note that $X_k, Y_k \in \mathcal{A}_k$. The quantity w_k denotes the weight assigned to the attribute A_k.

To understand how different measures calculate the per-attribute similarity, $S_k(X_k, Y_k)$, consider a categorical attribute A, which takes one of the values $\{a, b, c, d\}$. We have dropped the subscript k for simplicity. The per-attribute similarity computation is equivalent to constructing the (symmetric) matrix shown in Figure 1.

	a	b	c	d
a	$S(a,a)$	$S(a,b)$	$S(a,c)$	$S(a,d)$
b		$S(b,b)$	$S(b,c)$	$S(b,d)$
c			$S(c,c)$	$S(c,d)$
d				$S(d,d)$

Figure 1: Similarity Matrix for a Single Categorical Attribute

Essentially, in determining the similarity between two values, any categorical measure is filling the entries of this matrix. For example, the overlap measure sets the diagonal entries to 1 and the off-diagonal entries to 0, i.e., the similarity is 1 if the values match and 0 if the values mismatch. Additionally, measures may use the following information in computing a similarity value (all the measures in this paper use only this information):

- $f(a), f(b), f(c), f(d)$, the frequencies of the values in the data set
- N, the size of the data set
- n, the number of values taken by the attribute (4 in the case above)

We can classify measures in several ways, based on: (i) the manner in which they fill the entries of the similarity matrix, (ii) whether more weight is a function of the frequency of the attribute values, (iii) the arguments used to propose the measure (probabilistic, information-theoretic, etc.). In this paper, we will describe the measures by classifying them as follows:

- those that fill the *diagonal entries only*. These are measures that set the off-diagonal entries to 0 (mismatches are uniformly given the minimum value) and give possibly different weights to matches.
- those that fill the *off-diagonal entries only*. These measures set the diagonal entries to 1 (matches are uniformly given the maximum value) and give possibly different weights to mismatches.
- those that fill *both diagonal and off-diagonal* entries. These measures give different weights to both matches and mismatches.

Table 2 gives the mathematical formulas for the measures we will be describing in this paper. The various measures described in Table 2 compute the per-attribute similarity $S_k(X_k, Y_k)$ as shown in column 2 and compute the attribute weight w_k as shown in column 3.

4.1 Measures that fill Diagonal Entries only.

1. *Overlap*. The *overlap* measure simply counts the number of attributes that match in the two data instances. The range of per-attribute similarity for the *overlap* measure is $[0, 1]$, with a value of 0 occurring when there is no match, and a value of 1 occurring when the attribute values match.
2. *Goodall*. Goodall [14] proposed a measure that attempts to normalize the similarity between two objects by the probability that the similarity value observed could be observed in a random sample of two points. This measure assigns higher similarity to a match if the value is infrequent than if the value is frequent. Goodall's original measure details a procedure to combine similarities in the multivariate setting which takes into account dependencies between attributes. Since this procedure is computationally expensive, we use a simpler version of the measure (described next as *Goodall*1). Goodall's original measure is not empirically evaluated in this paper. We also propose three other variants of Goodall's measure in this paper: *Goodall*2, *Goodall*3 and *Goodall*4.

	Measure	$S_k(X_k, Y_k)$	$w_k,\ k = 1 \dots d$
1.	*Overlap*	$= \begin{cases} 1 & \text{if } X_k = Y_k \\ 0 & \text{otherwise} \end{cases}$	$\frac{1}{d}$
2.	*Eskin*	$= \begin{cases} 1 & \text{if } X_k = Y_k \\ \frac{n_k^2}{n_k^2+2} & \text{otherwise} \end{cases}$	$\frac{1}{d}$
3.	*IOF*	$= \begin{cases} 1 & \text{if } X_k = Y_k \\ \frac{1}{1+\log f_k(X_k) \times \log f_k(Y_k)} & \text{otherwise} \end{cases}$	$\frac{1}{d}$
4.	*OF*	$= \begin{cases} 1 & \text{if } X_k = Y_k \\ \frac{1}{1+\log \frac{N}{f_k(X_k)} \times \log \frac{N}{f_k(Y_k)}} & \text{otherwise} \end{cases}$	$\frac{1}{d}$
5.	*Lin*	$= \begin{cases} 2 \log \hat{p}_k(X_k) & \text{if } X_k = Y_k \\ 2 \log(\hat{p}_k(X_k) + \hat{p}_k(Y_k)) & \text{otherwise} \end{cases}$	$\frac{1}{\sum_{i=1}^{d} \log \hat{p}_i(X_i) + \log \hat{p}_i(Y_i)}$
6.	*Lin*1	$= \begin{cases} \sum_{q \in Q} \log \hat{p}_k(q) & \text{if } X_k = Y_k \\ 2 \log \sum_{q \in Q} \hat{p}_k(q) & \text{otherwise} \end{cases}$	$\frac{1}{\sum_{i=1}^{d} \sum_{q \in Q} \log \hat{p}_i(q)}$
7.	*Goodall*1	$= \begin{cases} 1 - \sum_{q \in Q} p_k^2(q) & \text{if } X_k = Y_k \\ 0 & \text{otherwise} \end{cases}$	$\frac{1}{d}$
8.	*Goodall*2	$= \begin{cases} 1 - \sum_{q \in Q} p_k^2(q) & \text{if } X_k = Y_k \\ 0 & \text{otherwise} \end{cases}$	$\frac{1}{d}$
9.	*Goodall*3	$= \begin{cases} 1 - p_k^2(X_k) & \text{if } X_k = Y_k \\ 0 & \text{otherwise} \end{cases}$	$\frac{1}{d}$
10.	*Goodall*4	$= \begin{cases} p_k^2(X_k) & \text{if } X_k = Y_k \\ 0 & \text{otherwise} \end{cases}$	$\frac{1}{d}$
11.	*Smirnov*	$= \begin{cases} 2 + \frac{N - f_k(X_k)}{f_k(X_k)} + \sum_{q \in \{\mathcal{A}_k \setminus X_k\}} \frac{f_k(q)}{N - f_k(q)} & \text{if } X_k = Y_k \\ \sum_{q \in \{\mathcal{A}_k \setminus \{X_k, Y_k\}\}} \frac{f_k(q)}{N - f_k(q)} & \text{otherwise} \end{cases}$	$\frac{1}{\sum_{k=1}^{d} n_k}$
12.	*Gambaryan*	$= \begin{cases} -[\hat{p}_k(X_k) \log_2 \hat{p}_k(X_k) + & \text{if } X_k = Y_k \\ (1 - \hat{p}_k(X_k)) \log_2 (1 - \hat{p}_k(X_k))] & \\ 0 & \text{otherwise} \end{cases}$	$\frac{1}{\sum_{k=1}^{d} n_k}$
13.	*Burnaby*	$= \begin{cases} 1 & \text{if } X_k = Y_k \\ \frac{\sum_{q \in \mathcal{A}_k} 2 \log(1 - \hat{p}_k(q))}{\log \frac{\hat{p}_k(X_k) \hat{p}_k(Y_k)}{(1 - \hat{p}_k(X_k))(1 - \hat{p}_k(Y_k))} + \sum_{q \in \mathcal{A}_k} 2 \log(1 - \hat{p}_k(q))} & \text{otherwise} \end{cases}$	$\frac{1}{d}$
14.	*Anderberg*	$S(X,Y) = \frac{\sum_{k \in \{1 \le k \le d : X_k = Y_k\}} \left(\frac{1}{\hat{p}_k(X_k)}\right)^2 \frac{2}{n_k(n_k+1)}}{\sum_{k \in \{1 \le k \le d : X_k = Y_k\}} \left(\frac{1}{\hat{p}_k(X_k)}\right)^2 \frac{2}{n_k(n_k+1)} + \sum_{k \in \{1 \le k \le d : X_k \ne Y_k\}} \left(\frac{1}{2\hat{p}_k(X_k)\hat{p}_k(Y_k)}\right) \frac{2}{n_k(n_k+1)}}$	

Table 2: Similarity Measures for Categorical Attributes. Note that $S(X,Y) = \sum_{k=1}^{d} w_k S_k(X_k, Y_k)$. For measure *Lin*1, $\{Q \subseteq \mathcal{A}_k : \forall q \in Q, \hat{p}_k(X_k) \le \hat{p}_k(q) \le \hat{p}_k(Y_k)\}$, assuming $\hat{p}_k(X_k) \le \hat{p}_k(Y_k)$. For measure *Goodall*1, $\{Q \subseteq \mathcal{A}_k : \forall q \in Q, p_k(q) \le p_k(X_k)\}$. For measure *Goodall*2, $\{Q \subseteq \mathcal{A}_k : \forall q \in Q, p_k(q) \ge p_k(X_k)\}$.

3. *Goodall*1. The *Goodall*1 measure is the same as Goodall's measure on a per-attribute basis. However, instead of combining the similarities by taking into account dependencies between attributes, the *Goodall*1 measure takes the average of the per-attribute similarities. The range of $S_k(X_k, Y_k)$ for matches in *Goodall*1 measure is $[0, 1 - \frac{2}{N(N-1)}]$, with the minimum being attained when attribute A_k takes only one value, and the maximum is attained when the value X_k occurs twice, while all other possible values of A_k occur more than 2 times.

4. *Goodall*2. The *Goodall*2 measure is a variant of Goodall's measure proposed by us. This measure assigns higher similarity if the matching values are infrequent, and at the same time there are other values that are even less frequent, i.e., the similarity is higher if there are many values with approximately equal frequencies, and lower if the frequency distribution is skewed. The range of $S_k(X_k, Y_k)$ for matches in the *Goodall*2 measure is $[0, 1 - \frac{2}{N(N-1)}]$, with the minimum value being attained if attribute A_k takes only one value, and the maximum is attained when the value X_k occurs twice, while all other possible values of A_k occur only once each.

5. *Goodall*3. We also propose another variant of Goodall's measure called *Goodall*3. The *Goodall*3 measure assigns a high similarity if the matching values are infrequent regardless of the frequencies of the other values. The range of $S_k(X_k, Y_k)$ for matches in the *Goodall*3 measure is $[0, 1 - \frac{2}{N(N-1)}]$, with the minimum value being attained if X_k is the only value for attribute A_k and maximum value is attained if X_k occurs only twice.

6. *Goodall*4. The *Goodall*4 measure assigns similarity $1 - Goodall3$ for matches. The range of $S_k(X_k, Y_k)$ for matches in the *Goodall*4 measure is $[\frac{2}{N(N-1)}, 1]$, with the minimum value being attained if X_k occurs only once, and the maximum value is attained if X_k is the only value for attribute A_k.

7. *Gambaryan*. Gambaryan proposed a measure [11] that gives more weight to matches where the matching value occurs in about half the data set, i.e., in between being frequent and rare. The *Gambaryan* measure for a single attribute match is closely related to the Shannon entropy from information theory, as can be seen from its formula in Table 2. The range of $S_k(X_k, Y_k)$ for matches in the *Gambaryan* measure is $[0, 1]$, with the minimum value being attained if X_k is the only value for attribute A_k and the maximum value is attained when X_k has frequency $\frac{N}{2}$.

4.2 Measures that fill Off-diagonal Entries only.

1. *Eskin*. Eskin et al. [9] proposed a normalization kernel for record-based network intrusion detection data. The original measure is distance-based and assigns a weight of $\frac{2}{n_k^2}$ for mismatches; when adapted to similarity, this becomes a weight of $\frac{n_k^2}{n_k^2+2}$. This measure gives more weight to mismatches that occur on attributes that take many values. The range of $S_k(X_k, Y_k)$ for mismatches in the *Eskin* measure is $[\frac{2}{3}, \frac{N^2}{N^2+2}]$, with the minimum value being attained when the attribute A_k takes only two values, and the maximum value is attained when the attribute has all unique values.

2. *Inverse Occurrence Frequency* (*IOF*). The inverse occurrence frequency measure assigns lower similarity to mismatches on more frequent values. The *IOF* measure is related to the concept of inverse document frequency which comes from information retrieval [19], where it is used to signify the relative number of documents that contain a specific word. A key difference is that inverse document frequency is computed on a term-document matrix which is usually binary, while the *IOF* measure is defined for categorical data. The range of $S_k(X_k, Y_k)$ for mismatches in the *IOF* measure is $[\frac{1}{1+(\log \frac{N}{2})^2}, 1]$, with the minimum value being attained when X_k and Y_k each occur $\frac{N}{2}$ times (i.e., these are the only two values), and the maximum value is attained when X_k and Y_k occur only once in the data set.

3. *Occurrence Frequency* (*OF*). The occurrence frequency measure gives the opposite weighting of the *IOF* measure for mismatches, i.e., mismatches on less frequent values are assigned lower similarity and mismatches on more frequent values are assigned higher similarity. The range of $S_k(X_k, Y_k)$ for mismatches in *OF* measure is $[\frac{1}{(1+(\log N)^2)}, \frac{1}{1+(\log 2)^2}]$, with the minimum value being attained when X_k and Y_k occur only once in the data set, and the maximum value attained when X_k and Y_k occur $\frac{N}{2}$ times.

4. *Burnaby*. Burnaby [5] proposed a similarity measure using arguments from information theory. He argues that the set of observed values are like a group of signals conveying information and, as in information theory, attribute values that are rarely observed should be considered more informative. In

[5], Burnaby proposed information weighted measures for binary, ordinal, categorical and continuous data. The measure we present in Table 2 is adapted from Burnaby's categorical measure. This measure assigns low similarity to mismatches on rare values and high similarity to mismatches on frequent values. The range of $S_k(X_k, Y_k)$ for mismatches for the *Burnaby* measure is $[\frac{N\log(1-\frac{1}{N})}{N\log(1-\frac{1}{N})-\log(N-1)}, 1]$, with the minimum value being attained when all values for attribute A_k occur only once, and the maximum value is attained when X_k and Y_k each occur $\frac{N}{2}$ times.

4.3 Measures that fill both Diagonal and Off-diagonal Entries.

1. *Lin.* In [23], Lin describes an information-theoretic framework for similarity, where he argues that when similarity is thought of in terms of assumptions about the space, the similarity measure naturally follows from the assumptions. Lin [23] discusses the ordinal, string, word and semantic similarity settings; we applied his framework to the categorical setting to derive the *Lin* measure in Table 2. The *Lin* measure gives higher weight to matches on frequent values, and lower weight to mismatches on infrequent values. The range of $S_k(X_k, Y_k)$ for a match in the *Lin* measure is $[-2\log N, 0]$, with the minimum value being attained when X_k occurs only once and the maximum value attained when X_k occurs N times. The range of $S_k(X_k, Y_k)$ for a mismatch in *Lin* measure is $[-2\log\frac{N}{2}, 0]$, with the minimum value being attained when X_k and Y_k each occur only once, and the maximum value is attained when X_k and Y_k each occur $\frac{N}{2}$ times.

2. *Lin*1. The *Lin*1 measure is another measure we have derived using Lin's similarity framework. This measure gives lower weight to mismatches if either of the mismatching values are very frequent, or if there are several values that have frequency in between those of the mismatching values; higher weight is given when there are mismatches on infrequent values and there are few other infrequent values. For matches, lower weight is given for matches on frequent values or matches on values that have many other values of the same frequency; higher weight is given to matches on rare values. The range of $S_k(X_k, Y_k)$ for matches in the *Lin*1 measure is $[-2\log\frac{N}{2}, 0]$, with the minimum value being attained when X_k occurs twice in the dataset, and no other value for attribute k occurs twice, and maximum value is attained when X_k occurs N times. The range of $S_k(X_k, Y_k)$ for mismatches in the *Lin*1 measure is $[-2\log\frac{N}{2}, 0]$, with the minimum value being attained when X_k and Y_k both occur only once and all other values for attribute A_k occur more than once, and the maximum value is attained when X_k is the most frequent value and Y_k is the least frequent value or vice versa.

3. *Smirnov.* Smirnov [31] proposed a measure rooted in probability theory that not only considers a given value's frequency, but also takes into account the distribution of the other values taken by the same attribute. The *Smirnov* measure is probabilistic for both matches and mismatches. For a match, the similarity is high when the frequency of the matching value is low, and the other values occur frequently. The range of $S_k(X_k, Y_k)$ for a match in the *Smirnov* measure is $[2, 2N]$, with the minimum value being attained when X_k occurs N times; the maximum value is attained when X_k occurs only once and there only one other possible value for attribute A_k, which occurs $N-1$ times. The range of $S_k(X_k, Y_k)$ for a mismatch in the *Smirnov* measure is $[0, \frac{N}{2}-1]$, with the minimum value being attained when the attribute A_k takes only two values, X_k and Y_k; and the maximum is attained when A_k takes only one more value apart from X_k and Y_k and it occurs $N-2$ times (X_k and Y_k occur once each).

4. *Anderberg.* In his book on cluster analysis [2], Anderberg presents an approach to handle similarity between categorical attributes. He argues that rare matches indicate a strong association and should be given a very high weight, and that mismatches on rare values should be treated as being distinctive and should also be given special importance. In accordance with these arguments, the *Anderberg* measure assigns higher similarity to rare matches, and lower similarity to rare mismatches. The *Anderberg* measure is unique in the sense that it cannot be written in the form of Equation 4.1. The range of the *Anderberg* measure is $[0, 1]$; the minimum value is attained when there are no matches, and the maximum value is attained when all attributes match.

4.4 Further classification of similarity measures. We can further classify categorical similarity measures based on the arguments used to proposed the measures:

1. *Probabilistic approaches* take into account the probability of a given match taking place. The following measures are probabilistic: $Goodall, Smirnov, Anderberg$.

2. *Information-theoretic approaches* incorporate the information content of a particular value/variable with respect to the data set. The following measures are information-theoretic: $Lin, Lin1, Burnaby$.

Table 3 provides a characterization of each of the 14 similarity measures in terms of how they handle the various characteristics of categorical data. This table shows that measures *Eskin* and *Anderberg* assign weight to every attribute using the quantity n_k, though in opposite ways. Another interesting observation from column 3 is that several measures—*Lin*, *Lin*1, *Goodall*1, *Goodall*3, *Smirnov*, *Anderberg*—assign higher similarity to a match when the attribute value is rare (f_k is low), while *Goodall*2 and *Goodall*4 assign higher similarity to a match when the attribute value is frequent (f_k is high). Only *Gambaryan* assigns the maximum similarity when the attribute value has a frequency close to $\frac{1}{2}$. Column 4 shows that *IOF*, *Lin*, *Lin*1, *Smirnov* and *Burnaby* assign greater similarity when the mismatch occurs between rare values, while *OF* and *Anderberg* assign greater similarity for a mismatch between frequent values.

5 Experimental Evaluation

In this section we present an experimental evaluation of the 14 measures (listed in Table 2) on 18 different data sets in the context of outlier detection.

Of these data sets, 16 are based on the data sets available at the UCI Machine Learning Repository [3], two are based on network data generated by SKAION Corporation for the ARDA information assurance program [30]. The details about the 18 data sets are summarized in Table 4. Eleven of these data sets were purely categorical, five (KD1,KD2,Sk1,Sk2,Cen) had a mix of continuous and categorical attributes, and two data sets, Irs and Sgm, were purely continuous. Continuous variables were discretized using the MDL method [10]. The KD1,KD2 data sets were obtained from the KDDCup data set by discretizing the continuous attributes into 10 and 100 bins respectively. 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.

Each data set contains labeled instances belonging to multiple classes. We identified one class as the outlier class, and rest of the classes were grouped together and called normal. For most of the data sets, the smallest class was selected as the outlier class. The only exceptions were (Cr1,Cr2) and (KD1,KD2), where the original data sets had two similar-sized classes. For each data set in the pair, we sampled 100 points from one of the two classes as the outlier points.

In Section 3.1 we discussed a number of characteristics for categorical data sets; in Table 4 we describe the various public data sets in terms of these characteristics. The first row gives the size of each data set. The second row shows the percentage of outlier points in the original data set. The third row indicates the number of attributes in the data sets. Rows 4 and 5 show the distribution of the number of values taken by each attribute; the difference between the average and the median is a measure of how skewed this distribution is. For example, data set Sk1 has a few attributes that take many values while most other attributes take few values. The next three rows show the distribution of frequency of values taken by an attribute in the given data set (i.e., $f_k(x)$). This is done by showing the number of attributes that have a uniform, Gaussian and skewed distribution in rows 6, 7, and 8 respectively.

The last two rows in Table 4 denote the cross-validation classification recall and precision reported by the C4.5 classifier on the outlier class. This quantity indicates the separability between the instances belonging to normal class(es) and instances belonging to outlier class, using the given set of attributes. A low accuracy implies that distinguishing between outliers and normal instances is difficult in that particular data set using a decision tree-based classifier.

5.1 Evaluation Methodology. The performance of the different similarity measures was evaluated in the context of outlier detection using nearest neighbors [29, 34]. All instances belonging to the normal class(es) form the training set. We construct the test set by adding the outlier points to the training set. For each test instance, we find its k nearest neighbors using the given similarity measure, in the training set (we chose the parameter $k = 10$). The outlier score is the distance to the k^{th} nearest neighbor. The test instances are then sorted in decreasing order of outlier scores.

To evaluate a measure, we count the number of true outliers in the top p portion of the sorted test instances, where $p = \delta n$, $0 \leq \delta \leq 1$ and 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 $\frac{o}{p}$.

In this paper we present results for $\delta = 1$. We have also experimented with other lower values of δ and the trends in relative performance are similar. We have presented these additional results in our extended work available as a technical report [6].

Measure	n_k	$\{f_k(X_k), f_k(Y_k)\}$	
		$X_k = Y_k$	$X_k \neq Y_k$
Overlap		1	0
Eskin	$\propto n_k^2$	1	0
IOF		1	$\propto 1/(\log f_k(X_k) \log f_k(Y_k))$
OF		1	$\propto \log f_k(X_k) \log f_k(Y_k)$
Lin		$\propto 1/\log f_k(X_k)$	$\propto 1/\log (f_k(X_k) + f_k(Y_k))$
*Lin*1		$\propto 1/\log f_k(X_k)$	$\propto 1/\log \lvert f_k(X_k) - f_k(Y_k)\rvert$
*Goodall*1		$\propto (1 - f_k^2(X_k))$	0
*Goodall*2		$\propto f_k^2(X_k)$	0
*Goodall*3		$\propto (1 - f_k^2(X_k))$	0
*Goodall*4		$\propto f_k^2(X_k)$	0
Smirnov		$\propto 1/f_k(X_k)$	$\propto 1/(f_k(X_k) + f_k(Y_k))$
Gambaryan		Maximum at $f_k(X_k) = \frac{N}{2}$	0
Burnaby		1	$\propto 1/\log f_k(X_k)$,$\propto 1/\log f_k(Y_k)$
Anderberg	$\propto 1/n_k$	$\propto 1/f_k^2(X_k)$	$\propto f_k(X_k) f_k(Y_k)$

Table 3: Relation between per-attribute similarity, $S(X_k, Y_k)$ and $\{n_k,\ f_k(X_k),\ f_k(Y_k)\}$.

5.2 Experimental Results on Public Data Sets. Our experimental results verified our initial hypotheses about categorical similarity measures. As can be seen from Table 5, there are many situations where the *Overlap* measure does not give good performance. This is consistent with our intuition that the use of additional information would lead to better performance. In particular, we expected that since categorical data does not have inherent ordering, data-driven measures would be able to take advantage of information present in the data set to make more accurate determinations of similarity between a pair of data instances.

We make some key observations about the results in Table 5:

1. No single measure is always superior or inferior. This is to be expected since each data set has different characteristics.

2. The use of some measures gives consistently better performance on a large variety of data. The *Lin*, *OF*, *Goodall*3 measures give among the best performance overall in terms of outlier detection performance. This is noteworthy since *Lin* and *Goodall*3 have been introduced for the first time in this paper.

3. There are some pairs of measures that exhibit complementary performance, i.e., one performs well where the other performs poorly and vice-versa. Example complimentary pairs are (OF, IOF), $(Lin, Lin1)$ and $(Goodall3, Goodall4)$. This observation means that it may be possible to construct measures that draw on the strengths of two measures in order to obtain superior performance. This is an aspect of this work that needs to be pursued in future work.

4. The performance of an outlier detection algorithm is significantly affected by the similarity measure used (we refer the reader to our extended work [6] for a similar evaluation using a different outlier detection algorithm, LOF, which provides similar conclusions). For example, for the *Cn* data set, which has a very low classification accuracy for the outlier class, using *OF* still achieves close to 50 % accuracy. We also note that for many of the data sets there is a relationship between decision tree performance (separability) and the performance of the measures. Specifically, for some of the data sets (e.g., Sk1, Tmr, Aud) with low separability there was high variance in the performance of the various measures.

5. The *Eskin* similarity measure weights attributes proportional to the number of values taken by the attribute (n_k). For data sets in which the attributes take large number of values (e.g., $KD2, Sk1, Sk2$), *Eskin* performs very poorly.

6. The *Smirnov* measure assigns similarity to both diagonal and off-diagonal entries in the per-attribute similarity matrix (Figure 1). But it still performs very poorly on most of the data sets. The other measures that operate similarly–*Lin*, *Lin*1 and *Anderberg*—perform better than *Smirnov* in almost every data set.

	Cr1	Cr2	Irs	Cn	KD1	KD2	Sk1	Sk2	Msh	Sgm	Cen	Can	Hys	Lym	Nur	Tmr	TTT	Aud
Size	1663	1659	150	4155	2100	2100	3480	2606	4308	210	5041	277	132	148	6480	336	727	190
% Outls.	4	4	33	2	5	5	4	4	3	14	16	29	23	41	3	33	14	25
d	6	6	4	42	29	29	10	10	21	18	10	9	4	18	8	15	9	17
avg(n_k)	3.23	3.28	6.80	2.91	4.80	37	337	286	4.18	6.37	7.09	3.90	2.80	3.10	3.33	2.31	2.80	2.65
med(n_k)	3	3	8	3	4	26	9	9	3	6	7	3	3	3	3	2	3	2
f_k Uni.	6	6	0	3	0	0	2	2	2	1	0	1	1	2	8	1	0	0
f_k Gauss.	0	0	4	7	8	7	3	3	5	8	7	5	3	11	0	2	9	5
f_k Skwd.	0	0	0	32	22	21	5	5	14	9	3	3	0	5	0	12	0	5
Recall	0.75	0.91	0.90	0.91	0.88	0.89	0.00	0.83	1.00	1.00	0.96	0.24	0.63	0.74	0.62	0.11	0.63	0.23
Precision	0.82	0.85	0.98	0.91	0.99	0.89	0.00	0.83	1.00	1.00	0.94	0.76	1.00	0.73	0.66	0.10	0.76	0.26

Table 4: Description of Public Data Sets.

Measure	Cr1	Cr2	Irs	Cn	KD1	KD2	Sk1	Sk2	Msh	Sgm	Cen	Can	Hys	Lym	Nur	Tmr	TTT	Aud	**Avg**
ovrlp	0.91	0.91	0.78	0.03	0.77	0.77	0.41	0.12	1.00	0.93	0.07	0.30	0.60	0.59	0.00	0.29	1.00	0.40	0.55
eskn	0.00	0.00	0.66	0.00	0.51	0.00	0.00	0.06	1.00	0.07	0.00	0.47	0.60	0.56	0.46	0.29	1.00	0.38	0.34
iof	1.00	0.92	0.20	0.40	0.44	0.10	0.10	0.10	1.00	0.00	0.15	0.52	0.60	0.56	0.49	0.36	1.00	0.36	0.46
of	0.93	0.92	0.88	0.48	0.74	0.90	0.57	0.36	1.00	1.00	0.27	0.36	1.00	0.69	0.47	0.22	0.28	0.62	0.65
lin	0.91	0.91	0.86	0.16	0.78	0.85	0.66	0.21	1.00	0.93	0.21	0.49	1.00	0.70	0.00	0.40	1.00	0.55	0.65
*lin*1	0.13	0.49	0.94	0.11	0.88	0.66	0.75	0.40	1.00	1.00	0.25	0.49	0.60	0.69	0.31	0.37	0.25	0.49	0.55
*good*1	0.70	0.68	0.80	0.11	0.77	0.01	0.59	0.40	1.00	1.00	0.36	0.30	0.87	0.72	0.00	0.47	0.75	0.40	0.55
*good*2	0.91	0.91	0.78	0.44	0.61	0.01	0.64	0.64	1.00	1.00	0.29	0.49	0.60	0.64	0.32	0.26	0.90	0.55	0.61
*good*3	0.91	0.91	0.80	0.16	0.75	0.01	0.59	0.41	1.00	1.00	0.37	0.46	0.87	0.67	0.04	0.50	1.00	0.43	0.60
*good*4	0.70	0.98	0.56	0.02	0.52	0.79	0.05	0.08	0.63	0.07	0.18	0.52	0.60	0.56	0.52	0.29	0.60	0.49	0.45
smrnv	0.91	0.91	0.78	0.03	0.00	0.00	0.32	0.04	0.00	0.83	0.15	0.23	0.70	0.38	0.08	0.33	0.55	0.04	0.35
gmbrn	0.91	0.91	0.76	0.47	0.43	0.01	0.14	0.35	1.00	0.97	0.25	0.56	0.60	0.70	0.46	0.35	1.00	0.60	0.58
brnby	0.91	0.91	0.78	0.03	0.66	0.90	0.55	0.17	1.00	0.93	0.13	0.51	0.90	0.59	0.46	0.29	1.00	0.45	0.62
anbrg	0.93	0.92	0.90	0.07	0.52	0.46	0.46	0.07	1.00	0.93	0.26	0.33	1.00	0.64	0.05	0.30	0.70	0.26	0.54
Avg	0.77	0.81	0.75	0.18	0.60	0.39	0.42	0.24	0.90	0.76	0.21	0.43	0.75	0.62	0.26	0.34	0.79	0.43	

Table 5: Experimental Results For kNN Algorithm for $\delta = 1.0$.

6 Concluding Remarks and Future Work

Computing similarity between categorical attributes has been discussed in a variety of contexts. In this paper we have brought together several such measures and evaluated them in the context of outlier detection. We have also proposed several variants ($Lin, Lin1, Goodall2, Goodall3, Goodall4$) of existing similarity measures, some of which perform very well as shown in our evaluation.

Given this set of similarity measures, the first question that comes to mind is: *Which similarity measure is best suited for my data mining task?*. Our experimental results suggest that there is no one best performing similarity measure. Hence, one needs to understand how a similarity measure handles the different characteristics of a categorical data set, and this needs to be explored in future research.

We used outlier detection as the underlying data mining task for the comparative evaluation in this work. However, similar studies can be performed using classification or clustering as the underlying task. It will be useful to know if the relative performance of these similarity measures remains the same for the other data mining tasks.

In our evaluation methodology we have used one similarity measure across all attributes. Since different attributes in a data set can be of different nature, an alternative way is to use different measures for different attributes. This appears to be especially promising given the complimentary nature of several similarity measures.

7 Acknowledgements

We are grateful to the anonymous reviewers for their comments and suggestions, which improved this paper. We would also like to thank György Simon for his helpful comments on an early draft of this paper. This work was supported by NSF Grant CNS-0551551, NSF ITR Grant ACI-0325949, NSF Grant IIS-0308264, and NSF Grant IIS-0713227. Access to computing facilities was provided by the University of Minnesota Digital Technology Center and Supercomputing Institute.

References

[1] A. Ahmad and L. Dey. A method to compute distance between two categorical values of same attribute in unsupervised learning for categorical data set. *Pattern Recogn. Lett.*, 28(1):110–118, 2007.

[2] M. R. Anderberg. *Cluster Analysis for Applications*. Academic Press, New York, 1973.

[3] A. Asuncion and D. J. Newman. UCI machine learning repository. [http://archive.ics.uci.edu/ml]. Irvine, CA: University of California, School of Information and Computer Science, 2007.

[4] Y. Biberman. A context similarity measure. In *ECML '94: Proceedings of the European Conference on Machine Learning*, pages 49–63. Springer, 1994.

[5] T. Burnaby. On a method for character weighting a similarity coefficient, employing the concept of information. *Mathematical Geology*, 2(1):25–38, 1970.

[6] V. Chandola, S. Boriah, and V. Kumar. Similarity measures for categorical data—a comparative study. Technical Report 07-022, Department of Computer Science & Engineering, University of Minnesota, October 2007.

[7] H. Cramér. *The Elements of Probability Theory and Some of its Applications*. John Wiley & Sons, New York, NY, 1946.

[8] G. Das and H. Mannila. Context-based similarity measures for categorical databases. In *PKDD '00: Proceedings of the 4th European Conference on Principles of Data Mining and Knowledge Discovery*, pages 201–210, London, UK, 2000. Springer-Verlag.

[9] E. Eskin, A. Arnold, M. Prerau, L. Portnoy, and S. Stolfo. A geometric framework for unsupervised anomaly detection. In D. Barbará and S. Jajodia, editors, *Applications of Data Mining in Computer Security*, pages 78–100. Kluwer Academic Publishers, Norwell, MA, 2002.

[10] 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.

[11] P. Gambaryan. A mathematical model of taxonomy. *Izvest. Akad. Nauk Armen. SSR*, 17(12):47–53, 1964.

[12] V. Ganti, J. Gehrke, and R. Ramakrishnan. CACTUS–clustering categorical data using summaries. In *KDD '99: Proceedings of the 5th 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] D. W. Goodall. A new similarity index based on probability. *Biometrics*, 22(4):882–907, 1966.

[15] S. Guha, R. Rastogi, and K. Shim. ROCK: A robust clustering algorithm for categorical attributes. *Information Systems*, 25(5):345–366, 2000.

[16] J. A. Hartigan. *Clustering Algorithms*. John Wiley & Sons, New York, NY, 1975.

[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] A. K. Jain and R. C. Dubes. *Algorithms for Clustering Data*. Prentice-Hall, Englewood Cliffs, NJ, 1988.

[19] K. S. Jones. A statistical interpretation of term specificity and its application in retrieval. In *Document Retrieval Systems*, volume 3 of *Taylor Graham Series In Foundations Of Information Science*, pages 132–142. Taylor Graham Publishing, London, UK, 1988. ISBN 0-947568-21-2.

[20] W. P. Jones and G. W. Furnas. Pictures of relevance: a geometric analysis of similarity measures. *J. Am. Soc. Inf. Sci.*, 38(6):420–442, 1987.

[21] L. Kaufman and P. J. Rousseeuw. *Finding Groups in Data: An Introduction to Cluster Analysis*. John Wiley & Sons, New York, NY, 1990.

[22] S. Q. Le and T. B. Ho. An association-based dissimilarity measure for categorical data. *Pattern Recogn. Lett.*, 26(16):2549–2557, 2005.

[23] D. Lin. An information-theoretic definition of similarity. In *ICML '98: Proceedings of the 15th International Conference on Machine Learning*, pages 296–304, San Francisco, CA, USA, 1998. Morgan Kaufmann Publishers Inc.

[24] K. Maung. Measurement of association in a contingency table with special reference to the pigmentation of hair and eye colours of Scottish school children. *Annals of Eugenics*, 11:189–223, 1941.

[25] T. Noreault, M. McGill, and M. B. Koll. A performance evaluation of similarity measures, document term weighting schemes and representations in a boolean environment. In *SIGIR '80: Proceedings of the 3rd annual ACM conference on Research and development in information retrieval*, pages 57–76, Kent, UK, 1981. Butterworth & Co.

[26] C. R. Palmer and C. Faloutsos. Electricity based external similarity of categorical attributes. In *PAKDD '03: Proceedings of the 7th Pacific-Asia Conference on Advances in Knowledge Discovery and Data Mining*, pages 486–500. Springer, 2003.

[27] C. P. Pappis and N. I. Karacapilidis. A comparative assessment of measures of similarity of fuzzy values. *Fuzzy Sets and Systems*, 56(2):171–174, 1993.

[28] K. Pearson. On the general theory of multiple contingency with special reference to partial contingency. *Biometrika*, 11(3):145–158, 1916.

[29] S. Ramaswamy, R. Rastogi, and K. Shim. Efficient algorithms for mining outliers from large data sets. In *SIGMOD '00: Proceedings of the ACM SIGMOD International Conference on Management of Data*, pages 427–438. ACM Press, 2000.

[30] SKAION Corporation. SKAION intrusion detection system evaluation data. [http://www.skaion.com/news/rel20031001.html].

[31] E. S. Smirnov. On exact methods in systematics. *Systematic Zoology*, 17(1):1–13, 1968.

[32] P. H. A. Sneath and R. R. Sokal. *Numerical Taxonomy: The Principles and Practice of Numerical Classification*. W. H. Freeman and Company, San Francisco, 1973.

[33] C. Stanfill and D. Waltz. Toward memory-based reasoning. *Commun. ACM*, 29(12):1213–1228, 1986.

[34] P.-N. Tan, M. Steinbach, and V. Kumar. *Introduction to Data Mining*. Addison-Wesley, Boston, MA, 2005.

[35] X. Wang, B. De Baets, and E. Kerre. A comparative study of similarity measures. *Fuzzy Sets and Systems*, 73(2):259–268, 1995.

[36] D. R. Wilson and T. R. Martinez. Improved heterogeneous distance functions. *J. Artif. Intell. Res. (JAIR)*, 6:1–34, 1997.

[37] R. Zwick, E. Carlstein, and D. V. Budescu. Measures of similarity among fuzzy concepts: A comparative analysis. *International Journal of Approximate Reasoning*, 1(2):221–242, 1987.