Information-theoretic and Set-theoretic Similarity

Luca Cazzanti
Applied Physics Lab
University of Washington
Seattle, WA 98195, USA
Email: luca@apl.washington.edu

Maya R. Gupta
Department of Electrical Engineering
University of Washington
Seattle, WA 98195, USA
Email: gupta@ee.washington.edu

***Abstract*— We introduce a definition of similarity based on Tversky's set-theoretic linear contrast model and on information-theoretic principles. The similarity measures the residual entropy with respect to a random object. This residual entropy similarity strongly captures context, which we conjecture is important for similarity-based statistical learning. Properties of the similarity definition are established and examples illustrate its characteristics. We show that a previously-defined information-theoretic similarity is also set-theoretic, and compare it to the residual entropy similarity. The similarity between random objects is also treated.**

I. Introduction

Similarity definitions are important for a range of classification, clustering, and other pattern recognition tasks. A recent review of the issues in assessing similarity for pattern recognition is given in [1]. Many similarity-based pattern recognition solutions have used application-specific notions of similarity. Here, we focus on definitions of similarity that can be applied across applications.

Set-theoretic definitions of similarity assume that each sample can be described as a set of features. This allows the application-specific information to be captured in the definition of relevant features, while employing a more general notion of similarity. A landmark model of set-theoretic similarity is Tversky's contrast similarity model [2]. Such general set-theoretic similarity definitions provide a basis for research into the theory and algorithms for similarity-based pattern-recognition.

Information theoretic ideas and analysis can be effective for assessing similarity for use in pattern recognition. Similarity measures based on information theory include information content [3], the amount of information needed to describe the commonality of two objects divided by the information needed to describe the objects [4], and similarity based on the (uncomputable) conditional Kolmogorov complexity of two objects [5].

In this paper, we first show that a previously defined information-theoretic similarity [4] is also a set-theoretic contrast similarity as per Tversky's definition. Then, we propose a set-theoretic definition of similarity that measures residual entropy in order to take into account context. We give a number of its properties and examples. Euclidean-based statistical learning theory generally assumes random training and test samples. The proposed residual entropy similarity provides natural notions of similarity for random objects.

II. Notation and Symbols

Let $\mathcal{O}$ be a finite set of objects over which one would like to define a similarity function. Let $\{R, A, B, E, F\} \in \mathcal{O}$ be independent and identically drawn random objects, such that a realization of a random object is an object, denoted by the corresponding lower-case letter, e.g. r. Suppose that a distribution exists over the set of objects $\mathcal{O}$ such that P is a well-defined probability mass function with $P(R = r) \geq 0$ and

$$\sum_{r \in \mathcal{O}} P(R = r) = 1.$$

Let each object be completely described by a finite set of features indexed by $i = 1, \ldots, n$ so that r_i is the ith feature of r. If $r_i = \emptyset$ then the ith feature of r is considered *missing* or *unspecified*. In general, $P(R = r) = P(R_1 = r_1, R_2 = r_2, \ldots, R_n = r_n)$, but unspecified features do not affect the probability, that is $P(R_1 = r_1, \ldots, R_{k-1} = r_{k-1}, R_k = \emptyset, R_{k+1} = r_{k+1}, \ldots R_n = r_n) = P(R_1 = r_1, \ldots, R_{k-1} = r_{k-1}, R_{k+1} = r_{k+1}, \ldots R_n = r_n)$. In some parts of the paper we will need to assume feature independence; in such cases $P(R = r) = \Pi_{i=1}^{n} P(R_i = r_i)$.

Let $a = b$ mean that a is an object that contains the same exact features as b. Let $a \subset r$ signify that all of the features in a are in object r. Then, the conditional probability of R given that $a \subset R$ is $P(R = r | a \subset R)$.

III. Set-theoretic Similarity

Research into similarity has a long history in psychology. An important contribution from that field is Tversky's set-theoretic similarity models [2], [6]. Tversky proposed considering objects as sets of features, and measuring the similarity between objects by appropriately weighting and combining the intersections and differences of the feature sets. His seminal *linear contrast model* for similarity between objects a and b is defined as

$$s(a, b) = \theta f(a \cap b) - \alpha f(a \setminus b) - \beta f(b \setminus a) \qquad (1)$$

where f is a positive *saliency* function that is monotonically increasing with respect to set inclusion, and θ, α and β are fixed positive real numbers. Thus a and b are more similar if their intersection increases, but less similar depending on which features belong exclusively to a or exclusively to b.

Tversky also proposed a *ratio* version of the contrast model where similarity is defined

$$s(a,b) = \frac{f(a \cap b)}{f(a \cap b) + \alpha f(a \setminus b) + \beta f(b \setminus a)}. \qquad (2)$$

Tversky's set-theoretic similarity models have been successful at explaining human similarity judgements in various similarity-assessment tasks, particularly when the objects are not described well by low-level numerical features, and when the assessment of similarity involves considerable cognitive effort [7], [8], [9]. Tversky's models of similarity have been applied outside of psychology. One example is the use of Tversky's model of similarity for searching structural chemical information databases [10]. It can be easily shown that Tversky's contrast models generalize the Hamming, Jaccard, and other distance measures commonly used to assess similarity between binary vectors [11], [10]. One simply describes sets as binary vectors in which a 1 (or 0) indicates the presence (or absence) of a feature. The saliency function is the sum of the elements of a binary vector, or, more generally, the cardinality of a set.

Tversky's similarity contrast models do not generally satisfy the properties for a metric (minimality, symmetry, and the triangle inequality), and Tversky presents evidence from experiments where humans assess similarity to show that these properties are not necessarily correlated to how humans judge similarity [2], [6].

IV. Information-theoretic Similarity

In this section, Lin's information-theoretic similarity [4] is shown to be an example of a contrast set-theoretic similarity. Then, Resnik's information-theoretic similarity [3], [12] is reviewed, and shown not to follow the contrast set-theoretic model. Another general information-theoretic similarity is offered by Li et al. [5]; they consider the class of metric similarities for sequences and explain why the conditional Kolmogorov complexity is optimal for measuring similarity. The focus of this paper is more general, in that an information-theoretic similarity is sought for arbitrary objects that can be defined as sets of features, which includes the class of sequences.

Lin [4] defines an information-theoretic similarity measure based on the information content of feature vectors, where information of each feature is defined in the standard way $I(r) = -\log P(r)$:

$$s_{Lin}(a,b) = \frac{2I(\texttt{common}(a,b))}{I(\texttt{description}(a,b))}.$$

By $\texttt{common}(a,b)$, Lin means the set of features common to both objects a and b; by $\texttt{description}(a,b)$ he means the set of features needed to completely describe both objects a and b. Lin assumes that features are independent in his examples.

Lemma 1: Let the features be probabilistically independent, then Lin's similarity measure is an example of Tversky's ratio contrast similarity as per (2), with saliency function $f = I$, and $\alpha = \beta = 0.5$.

Proof: We use feature independence and the properties of the logarithm to break the commonality and the descriptions of a and b into subsets of features in order to arrive at Tversky's ratio contrast model. The numerator is simply $I(\texttt{common}(a,b)) = I(a \cap b)$. The denominator may be written as $I(\texttt{description}(a,b)) = I(a) + I(b) = 2I(a \cap b) + I(a \setminus b) + I(b \setminus a)$. Lin's similarity can thus be written as a ratio contrast similarity with saliency function $f = I$:

$$s_{Lin} = \frac{I(a \cap b)}{I(a \cap b) + 0.5I(a \setminus b) + 0.5I(b \setminus a)}.$$

□

Resnik [3], [12] introduces a measure of similarity for concepts in a IS-A taxonomy based on the information content of their parent concepts. For example, both "cash" and "credit" are children of the parent concept "medium of exchange"; both "coin" and "bill" are subconcepts of "cash". Resnik defines the similarity between any two children subconcepts c_1 and c_2 as the maximum information content evaluated over the set C of all parent concepts which subsume c_1 and c_2,

$$s_{Resnik}(c_1, c_2) = \max_{c \in C}[-\log P(c)],$$

where $P(c)$ is the empirical probability of concept c calculated from a dataset. Thus, the similarity $s_{Resnik}(c1, c2)$ is the information associated with the most improbable concept that includes the union of subconcepts c_1 and c_2.

If one represents each concept as a set of features, then Resnik's similarity is the maximally informative feature set out of all parent concepts that subsume c_1 and c_2. One expects the intersection of the features of c_1 and c_2 to be features of its direct parent concept. But the parent concept c^* which yields the maximum information may have a feature set very different from that of c_1 and c_2. Thus, Resnik's similarity can be interpreted in terms of sets of features, but is not a set-theoretic contrast similarity in that $s_{Resnick}(c_1, c_2)$ is not a positive function of common elements of c_1 and c_2 and a negative function of distinct elements of c_1 and c_2.

V. Context-dependent Similarity

Context is an important component of determining the similarity of objects [2], [6]. For example, given a set of coins, two Egyptian coins might be considered more similar if no other coins are Egyptian coins than if most of the coins are Egyptian coins. Context may play a key role in similarity-based pattern recognition because collected training samples for a statistical learning task are not arbitrary: informative features imply that there is relevant information about the class labels in the sample space of training data. For example, consider two classification tasks: classifying incoming emails as spam or not spam, and classifying saved emails as work or personal. In the two different contexts, it would aid the classification if the context were automatically taken into account when calculating the similarity. For the spam and not spam problem, the similarity between an email from a friend and from a colleague should be relatively higher than the similarity between those same emails for the problem of classifying into personal and work emails. By defining similarity within the context of the set of training samples being used, it may be possible to capture this important contextual information and improve statistical learning accuracy.

Lin's similarity takes into account context by incorporating the probability of features (their "information") into the similarity definition. The similarity is greater if the common features are less likely. For some applications, the context will be very important and must be captured more strongly by the similarity function. To this end, one might ask, if one knew that a random object R was at least described by the features in common between objects a and b, then how uncertain would one still be about R? This would specify the similarity of a and b in light of the context of the distribution of R. Next, we define a similarity based on this idea, present some of its properties, and illustrate how it more strongly captures context than Lin's similarity.

Let the *residual entropy similarity* be a Tversky linear contrast model as per (1) with $\theta = 1$, $\alpha = \beta = 0.5$, and the saliency function is the mutual information: $f(a \cap b) = I(R; a \cap b \subset R)$, $f(a \setminus b) = I(R; a \setminus b \subset R)$, and $f(b \setminus a) = I(R; b \setminus a \subset R)$. Then, because $I(R;X) = H(R) - H(R|X)$, the residual entropy similarity is defined:

$$\begin{aligned} s_{re}(a,b) = &-H(R|a \cap b \subset R) \\ &+\frac{H(R|a \setminus b \subset R)}{2} + \frac{H(R|b \setminus a \subset R)}{2}. \end{aligned} \tag{3}$$

where $H(R|a \cap b \in R) = -\sum_{r \in \mathcal{O}} P(R = r|a \cap b \in R) \log P(R = r|a \cap b \in R)$. If one makes the assumption that the features are probabilistically independent,

$$\begin{aligned} &H(R|a \cap b \in R) = \\ &\quad -\sum_{r \in \mathcal{O}} \Pi_{i=1}^{n} P(R_i = r_i) \log(P(R_i = r_i) \mathbf{1}_{((a \cap b)_i \neq \emptyset)}, \end{aligned}$$

where $\mathbf{1}_{()}$ is the indicator function.

Note that the residual entropy similarity $s_{re}(a,b)$ is a function of the object space $\mathcal{O}$ and of the distribution P of random object R; more explicit notation would be $s_{re}(a,b,P)$. The similarity s_{re} is only well-defined if $P(a \cap b \subset R) \neq 0$. In fact, in the notation section A, B, R were defined to be iid, so that a and b are realizations of the random objects A and B and thus $P(a \cap b \subset R) \neq 0$. More generally, the similarity $s_{re}(x,z)$ is well-defined if: x and z are realizations of random objects X and Z distributed respectively with P_X and P_Z such that P_X and P_Z are both absolutely continuous with respect to P.

Next, we establish some properties of s_{re}.

A. Properties of residual entropy similarity

1) Assume that object a is fixed, then $s_{re}(a,b)$ is maximum for $b = a$ and $s_{re}(a,a) = H(R) - H(R|a \subset R) = I(R; a \subset R)$. To maximize $s_{re}(a,b)$ over b one seeks to increase the size of the commonality (intersection) between a and b and decrease their differences (set exclusions). These goals are optimized by setting $b = a$.

2) The residual entropy function does not obey the *minimality* property of distances, that is $s_{re}(a,a) \neq s_{re}(b,b)$. Experimental tests of how people judge similarity have shown that minimality often does not hold [2]. As an example, given a set of coins, it is reasonable and may be useful to define the similarity between the 534th penny and itself to be smaller than the similarity between an ancient Egyptian coin and itself.

3) The maximum residual entropy similarity is $H(R)$ and it is achieved when $a \cap b$ uniquely specifies a random object (so that $H(R|a \cap b \subset R) = 0$), and the probability of the exclusions $a \setminus b$ and $b \setminus a$ are one: $P(a \setminus b) = 1$ so that $H(R|a \setminus b \subset R) = H(R)$. But if the probability of the exclusions is one, then the sets $a \setminus b$ and $b \setminus a$ must be empty, because a and b are realizations of random objects drawn from the same distribution. Since $a \setminus b = \emptyset$ and $b \setminus a = \emptyset$, it must be that $a = b$. Thus, the maximum similarity is the self-similarity $s_{re}(a,a)$ for some object a that has a feature set that uniquely describes a random object (that is, $H(R|a \subset R) = 0$).

4) The minimum similarity is $-H(R)$ and it is achieved if $a \cap b = \emptyset$ and a and b each uniquely specify R. Under these conditions, $H(R|a \cap b \subset R) = H(R)$ and $a \setminus b = a$, so that $H(R|a \setminus b \subset R) = H(R|a \subset R) = 0$ and $H(R|b \setminus a \subset R) = H(R|b \subset R) = 0$.

5) Consider two different distributions P_{R_1} and P_{R_2} over the same set of objects $\mathcal{O}$. The range of similarities with respect to these two different distributions may be different, with minimum similarities $-H(R_1)$ and $-H(R_2)$ and maximum similarities $H(R_1)$ and $H(R_2)$.

6) Symmetry holds, such that for any two objects a and b, $s_{re}(a,b) = s_{re}(b,a)$. Symmetry follows from $\alpha = \beta$. Psychologists have found that such symmetry does not always hold for human similarity judgements. Our residual entropy similarity can flexibly accommodate asymmetrical similarity relations by appropriately setting $\alpha \neq \beta$.

7) The triangle inequality does not hold. For similarity, the triangle inequality is written as $s(a,c) \geq s(a,b) + s(b,c)$, that is any two objects in a triplet are at least as similar as the sum of their similarities to the third object. Lin [4] and Tversky [2], [6] provide intuitive counterexamples which show that in general similarity relations are not transitive and do not obey the triangle inequality. Lin's counterexample for s_{Lin} (repeated in Figure 1) also holds for s_{re}. In this case there are two features

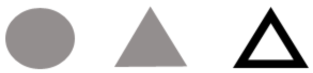

Fig. 1. Counterexample of triangle inequality.

"shape" and "shading", and each feature can take one of two values, "circle" or "triangle" and "white" or "gray". Numerically, the empirical probabilities associated with the features are $P(circle) = 1/3$ and $P(white) = 1/3$, which give $s_{re}(a, b) = s_{re}(b, c) = 0$ but $s_{re}(a, c) = -H(R)$, and thus the triangle inequality does not hold. Note that $a \cap b = \emptyset$, thus their similarity is the minimum achievable value.

8) Monotonicity holds, as defined by Tversky [2] such that $s(a, b) \geq s(a, c)$ whenever $(a \cap b) \subset (a \cap c)$ and $(a \setminus b) \subset (a \setminus c)$, because increased conditioning can only reduce entropy.

VI. Comparison of similarities that are information-theoretic and set-theoretic

Lin's similarity measures how improbable the features of the intersection of two objects are, normalized by the improbability of each object's features. The residual entropy similarity function measures the *expected* uncertainty given the intersection of the two objects, minus the expected uncertainty given their exclusions. The following examples help illustrate the two similarity functions:

Example 1: Consider two objects a and b. The self-similarities $s_{Lin}(a, a) = s_{Lin}(b, b) = 1$. In contrast, $s_{re}(a, a)$ may not equal $s_{re}(b, b)$, and both self-similarities are upperbounded by $H(R)$. In general, $s_{re}(a, a)$ is higher if a is less probable than b is.

Example 2: Consider two object pairs a, b and e, f such that $P(a \cap b \subset R) = 1/10 = P(e \cap f \subset R)$. Let $R = e \cap f$ if $e \cap f \subset R$. In contrast, given $a \cap b \subset R$, let $R = r_1$ with probability $1/2$, and $R = r_2$ with probability $1/2$. Then $H(R|e \cap f \subset R) = 0 < H(R|a \cap b \subset R)$. That is, $e \cap f$ more uniquely specifies an object than $a \cap b$. Let the other terms of $s_{re}(a, b)$ and $s_{Lin}(a, b)$ be equivalent, that is let $\sum_i \log P(a_i) + \log P(b_i) = \sum_i \log P(e_i) + \log P(f_i)$ and $H(R|a \setminus b) + H(R|b \setminus a) = H(R|e \setminus f) + H(R|f \setminus e)$. Then $s_{Lin}(a, b) = s_{Lin}(e, f)$, but $s_{re}(a, b) < s_{re}(e, f)$. Lin's similarity treats these two object pairs as equivalent, but the residual entropy similarity views e and f as more similar because the intersection of e and f more uniquely specifies an object.

Example 3: Consider two pairs of objects, (a, b) and (e, f), such that $P(a \cap b \subset R) = 1/5$ and $P(e \cap f \subset R) = 1/10$. Let $H(R|a \cap b \subset R) = H(R|e \cap f \subset R)$. For example, if $a \cap b$ uniquely specifies R and $e \cap f$ uniquely specifies R, then $H(R|a \cap b \subset R) = H(R|e \cap f \subset R) = 0$. Let the other terms of $s_{re}(a, b)$ and $s_{Lin}(a, b)$ be equivalent, that is, let $\sum_i \log P(a_i) + \log P(b_i) = \sum_i \log P(e_i) + P(f_i)$ and $H(R|a \setminus b \subset R) + H(R|b \setminus a \subset R) = H(R|e \setminus f \subset R) + H(R|f \setminus e \subset R)$.

Then $s_{Lin}(a, b) < s_{Lin}(e, f)$ because the intersection of e and f is less probable. In contrast, $s_{re}(a, b) = s_{re}(e, f)$ because given either intersection the uncertainty is the same.

Example 4: Consider two objects a and b such that $a \cap b = \emptyset$. Then $s_{Lin}(a, b) = 0$. In contrast, $s_{re}(a, b)$ will depend on how well the exclusions $a \setminus b = a$ and $b \setminus a = b$ specify R, and $s_{re}(a, b)$ achieves the minimum similarity $-H(R)$ if and only if a and b both uniquely specify R.

VII. Similarity of Random Objects

The similarity of random objects can be compared using the residual entropy similarity. As an example suppose one had a set of coins, and X is a random coin that is an Egyptian coin with probability $1/3$ or a British coin with probability $2/3$, and Z is a random coin that is made of copper with probability $5/7$, of silver with probability $1/7$, and of gold with probability $1/7$. Then, one can measure the similarity of X and Z as a positive function of how well the probabilistic intersection of X and Z specify a random coin, and a negative function of how well the probabilistic exclusions of X and Z specify a coin. Note that the distribution P over the object set $\mathcal{O}$ defines the context in which X and Z are being compared. This similarity of two random objects is a real number $s_{re}(X, Z)$.

Next we consider the random similarity $S_{re}(X, Z)$ between two random objects X and Z. That is, $S_{re}(X, Z)$ is a random variable that takes on value $s_{re}(x, z)$ with probability $P(X = x, Z = z)$. In the next paragraphs we formalize these two notions of the similarity of random objects, $s_{re}(X, Z)$ and $S_{re}(X, Z)$, and show that they are related in that $E[S_{re}(X, Z)] = s_{re}(X, Z)$.

Let $X, Z \in \mathcal{O}$ be independent random objects with probability distributions P_X and P_Z such that P_X and P_Z are absolutely continuous with respect to P. Then $X \cap Z$ is a random object with probability $P(X \cap Z = \phi)$ for any object $\phi \in \mathcal{O}$. Define $X \setminus Z$ and $Z \setminus X$ in the same manner.

Then the deterministic residual entropy similarity of random objects X and Z is

$$\begin{aligned} s_{re}(X, Z) &= -H(R|X \cap Z \subset R) \\ &+ \frac{H(R|X \setminus Z \subset R)}{2} + \frac{H(R|Z \setminus X \subset R)}{2}, \end{aligned} \tag{4}$$

where

$$H(R|X \cap Z \subset R) = \sum_{\phi \in \mathcal{O}} P(X \cap Z = \phi) H(R|\phi \subset R). \tag{5}$$

Also, a random similarity of random objects X and Z can be defined where $S_{re}(X, Z) = s_{re}(x, z)$ with $P(X = x, Z = z)$.

Then the expected random similarity can be defined

$$E[S_{re}(X,Z)] = \sum_{x \in \mathcal{O}} \sum_{z \in \mathcal{O}} P(X=x, Z=z) s_{re}(x,z). \quad (6)$$

There exists a well-developed general theory of random sets [13], which considers random sets that have realizations that may be continuous and infinite (for example, the set defined by an open line segment on the real line). In that general theory there are a number of different notions of expectation. For this work, the sets are constrained to be finite sets of discrete features selected from a finite discrete sample space $\mathcal{O}$, and thus much of the complex machinery needed for general random set theory is not needed here.

Lemma 2: $E[S_{re}(X,Z)] = s_{re}(X,Z)$.

Proof: The similarity $s_{re}(X,Z)$ is composed of three terms, and $E[S_{re}(X,Z)]$ can be similarly expanded into three terms by expanding the term $s_{re}(X,Z)$ in (6) into its three terms as per (3) and distributing. Then the first term of $E[S_{re}(X,Z)]$ is

$$\begin{aligned}
&\sum_{x \in \mathcal{O}} \sum_{z \in \mathcal{O}} P(X=x, Z=z) H(R|x \cap z \subset R) \\
&= \sum_{x \in \mathcal{O}} \sum_{z \in \mathcal{O}} P(X=x, Z=z) H(R|x \cap z \subset R)(1) \\
&= \sum_{x \in \mathcal{O}} \sum_{z \in \mathcal{O}} P(X=x, Z=z) H(R|x \cap z \subset R) \\
&\left(\sum_{\phi \in \mathcal{O}} P(X \cap Z = \phi)\right) \\
&= \sum_{\phi \in \mathcal{O}} \sum_{x \in \mathcal{O}} \sum_{z \in \mathcal{O}} P(X=x, Z=z) \\
&P(X \cap Z = \phi) H(R|\phi \subset R) \\
&= \sum_{\phi \in \mathcal{O}} P(X \cap Z = \phi) H(R|\phi \subset R) \\
&\left(\sum_{x \in \mathcal{O}} \sum_{z \in \mathcal{O}} P(X=x, Z=z)\right) \\
&= \sum_{\phi \in \mathcal{O}} P(X \cap Z = \phi) H(R|\phi \subset R) \qquad (7)
\end{aligned}$$

It is seen that (7) is equivalent to the first term of $s(X,Z)$, as per (4) and (5). The same logic holds for the second and third term. Thus, every term of $E[S_{re}(X,Z)]$ is equal to every term of $s_{re}(X,Z)$. □

VIII. Discussion

The proposed s_{re} is a set-theoretic definition of similarity that uses information theory to take into account context when assessing the similarity of two objects. The similarity s_{re} has different properties than other proposed similarity functions, and its properties reflect the findings of Tversky and others about human conceptions of similarity.

Theoretical results for similarity-based learning can be based on drawing iid training and test samples from class conditional distributions, as is often assumed for the theory of standard Euclidean-based learning [14]. Consider an example of similarity-based statistical learning. Suppose a set of objects $\{x_j\}$, $j = 1, \ldots, J$ is drawn from the distribution of spam emails with pmf P_X, and suppose another set of objects $\{z_k\}$, $k = 1, \ldots, K$ is drawn from the distribution of not spam emails P_Z. A "test" email t is drawn from the distribution $P_R = .5P_X + .5P_Z$. The similarities $\{s_{re}(t, x_j, P_R), s_{re}(t, z_k, P_R)\}$ are well-defined because P_X and P_Z are absolutely continuous with respect to P_R. Then, a 1-nearest neighbor decision rule would classify t as the class label of the training object that maximizes the similarity s_{re}.

In general, the development of similarity-based learning theory and algorithms may benefit from similarity functions that are set-theoretic, information-theoretic, that capture context, and that are well-defined for random objects from different class conditional distributions.

Acknowledgment

This work was funded in part by the Office of Naval Research, Code 321, Grant # N00014-05-1-0843. The authors thank Warren Fox, Eric Garcia, John Miyamoto, and Santosh Srivastava for helpful discussions.

References

[1] S. Santini and R. Jain, "Similarity measures," *IEEE Trans. Pattern Anal. Machine Intell.*, vol. 21, no. 9, pp. 871–883, September 1999.

[2] A. Tversky, "Features of similarity," *Psychological Review*, no. 84, pp. 327–352, 1977.

[3] P. Resnik, "Using information content to evaluate semantic similarity in a taxonomy," *Proc. IJCAI*, pp. 448–453, 1995.

[4] D. Lin, "An information-theoretic definition of similarity," *Proc. of the Intl. Conf. on Machine Learning*, 1998.

[5] M. Li, X. Chen, X. Li, B. Ma, and P. M. B. Vitányi, "The similarity metric," *IEEE Trans. Pattern Anal. Machine Intell.*, vol. 50, no. 12, pp. 3250–3264, December 2004.

[6] A. Tversky and I. Gati, "Studies of similarity," in *Cognition and Categorization*, E. Rosch and B. Lloyd, Eds. Hillsdale, N.J.: Earlbaum, 1978.

[7] I. Gati and A. Tversky, "Weighting common and distinctive features in perceptual and conceptual judgments," *Cognitive Psychology*, no. 16, pp. 341–370, 1984.

[8] A. Tversky and J. W. Hutchinson, "Nearest neighbor analysis of psychological spaces," *Psychological Review*, vol. 93, pp. 3–22, 1986.

[9] S. Sattath and A. Tversky, "On the relation between common and distinctive feature models," *Psychological Review*, no. 94, pp. 16–22, 1987.

[10] P. Willet, J. M. Barnard, and G. M. Downs, "Chemical similarity searching," *J. Chem. Inf. Comput. Sci.*, vol. 38, no. 6, pp. 983–996, 1998.

[11] B. Zhang and S. Shrihari, "Discovery of the tri-edge inequality with several binary dissimilarity measures," *Proc. of the Intl. Conf. on Pattern Recognition*, vol. 4, pp. 669–672, August 2004.

[12] P. Resnik, "Semantic similarity in a taxonomy: An information-based measure and its application to problems of ambiguity in natural language," *J. Art. Intell. Res.*, July 1999.

[13] I. Molchanov, *Theory of Random Sets*. London: Springer-Verlag, 2005.

[14] L. Devroye, L. Gyorfi, and G. Lugosi, *A Probabilistic Theory of Pattern Recognition*. New York: Springer-Verlag Inc., 1996.