Normalized Compression Distance of Multiples[1]

Andrew R. Cohen and Paul M.B. Vitányi

arXiv:1212.5711v1 [cs.CV] 22 Dec 2012

Abstract

Normalized compression distance (NCD) is a parameter-free similarity measure based on compression. The NCD between pairs of objects is not sufficient for all applications. We propose an NCD of finite multisets (multiples) of objacts that is metric and is better for many applications. Previously, attempts to obtain such an NCD failed. We use the theoretical notion of Kolmogorov complexity that for practical purposes is approximated from above by the length of the compressed version of the file involved, using a real-world compression program. We applied the new NCD for multiples to retinal progenitor cell questions that were earlier treated with the pairwise NCD. Here we get significantly better results. We also applied the NCD for multiples to synthetic time sequence data. The preliminary results are as good as nearest neighbor Euclidean classifier.

Index Terms— Normalized compression distance, multisets or multiples, pattern recognition, data mining, similarity, Kolmogorov complexity, retinal progenitor cell classification, synthetic data classification

I. Introduction

The classical notion of Kolmogorov complexity [13] is an objective measure for the information in an a *single* object, and information distance measures the information between a *pair* of objects [2]. This last notion has spawned research in the theoretical direction, for example [21], and in the practical direction through the *normalized* compression distance, the similarity metric, which arises by normalizing the information distance in a proper manner and approximating the Kolmogorov complexity through real-world compressors [16], [4], This normalized compression distance is a parameter-free, feature-free, and alignment-free similarity measure that has found many applications in pattern recognition, phylogeny,

Andrew Cohen is with the Department of Electrical and Computer Engineering, Drexel University. Address: A.R. Cohen, 3120–40 Market Street, Suite 313, Philadelphia, PA 19104, USA. Email: acohen@coe.drexel.edu

Paul Vitányi is with the national research center for mathematics and computer science in the Netherlands (CWI), and the University of Amsterdam. Address: CWI, Science Park 123, 1098XG Amsterdam, The Netherlands. Email: Paul.Vitanyi@cwi.nl.

clustering, and classification, for example [1], [11], [12], [22], [23], [6], [7], [27]. Another application is to objects that are only represented by name, or objects that are abstract like 'red,' 'Einstein,' 'three.' In this case the similarity metric uses background information provided by Google or any search engine that produces aggregate page counts. It discovers the 'meaning' of words and phrases in the sense of producing a relative semantics [5].

However, in many applications we are interested in shared information between many objects instead of just a pair of objects. In customer reviews of gadgets, in blogs about public happenings, in newspaper articles about the same occurrence, we are interested in the most comprehensive one or the most specialized one. Thus, the information distance measure has been extended from pairs to finite multisets. For many applications we require a normalized and computable version. For instance, classifying an object into one or another of disjoint classes we aim for the class of which the NCD for multiples grows the least. We applied the new NCD for multiples to retinal progenitor cell questions that were earlier treated with the pairwise NCD. Here we get significantly better results. We also applied the NCD for multiples to synthetic time sequence data. The preliminary results are as good as nearest neighbor Euclidean classifier.

A. Related Work

In [17] the notion is introduced of the information required to go from any object in a multiset of objects to any other object in the multiset. This is applied to extracting the essence from, for example, a finite multiset of internet news items, reviews of electronic cameras, tv's, and so on, in a way that works better than other methods. Let X denote a finite multiset of m finite binary strings defined by (abusing the set notation) $X = \{x_1, \ldots, x_m\}$, the constituting elements (not necessarily all different) ordered length-increasing lexicographic. We use multisets and not sets, since if X is a set then all of its members are different while we are interested in the situation were some of the objects are equal. Let U be the reference universal Turing machine, for convenience the prefix one as in Section II. We define the *information distance* in X by $E_{\max}(X) = \min\{|p| : U(x_i, p, j) = x_j \text{ for all } x_i, x_j \in X\}$. It is shown in [17], Theorem 2, that

$$E_{\max}(X) = \max_{x:x\in X} K(X|x), \tag{I.1}$$

up to a logarithmic additive term. Define $E_{\min}(X) = \min_{x:x\in X} K(X|x)$. Theorem 3 in [17] states that

$$E_{\min}(X) \leq E_{\max}(X) \leq \min_{i:1\leq i\leq m} \sum_{x_i,x_k\in X \ \& \ k\neq i} E_{\max}(x_i, x_k), \tag{I.2}$$

up to a logarithmic additive term. The paper [17] develops the stated results and applications. The information distance in [2] between strings x_1 and x_2 is denoted $E_{\max}(x_1, x_2) = \max\{K(x_1|x_2), K(x_2|x_1)\}$. Here we use the notation $E_{\max}(X) = \max_{x:x\in X} K(X|x)$. The two coincide for $|X| = 2$ since $K(x, y|x) = K(y|x)$ up to an additive constant term. In [24] the following results were obtained for multisets. The maximal overlap of information, concerning the remarkable property that the information needed to go from any member x_j to any other member x_k in a multiset X can be divided in two parts: a single string of length $\min_i K(X|x_i)$ and a special string of length $\max_i(K(X|x_i) - \min_i K(X|x_i)$ possibly depending on j and some logarithmic additive terms possibly depending on j, k. Furthermore, the minimal overlap property. the metricity property, the universality property, and the not-subadditivity property. With respect to normalization of the information distance of multisets abortive attempts were given. A review of some of the above is [18].

II. Preliminaries

A. Kolmogorov Complexity

The Kolmogorov complexity is the information in a single object [13]. Informally, the Kolmogorov complexity of a finite binary string is the length of the shortest string from which the original can be losslessly reconstructed by an effective general-purpose computer such as a particular universal Turing machine. Hence it constitutes a lower bound on how far a lossless compression program can compress. For technical reasons we choose Turing machines with a separate read-only input tape that is scanned from left to right without backing up, a separate work tape on which the computation takes place, and a separate output tape. All tapes are divided into squares and are sem-infinite. Initially the input tape contains a semi-infinite binary string with one bit per square starting at the leftmost square. Upon halting, the initial segment p of the input that has been scanned is called the input "program" and the contents of the output tape is called the "output." By construction, the set of halting programs is prefix free. We call U the reference universal prefix Turing machine. This leads to the definition of "prefix Kolmogorov complexity" which we shall designate simply as "Kolmogorov complexity."

Formally, the *conditional Kolmogorov complexity* $K(x|y)$ is the length of the shortest input z such that the reference universal prefix Turing machine U on input z with auxiliary information y outputs x. The

unconditional Kolmogorov complexity $K(x)$ is defined by $K(x|\epsilon)$ where ϵ is the empty string (of length 0). In these definitions both x and y can consist of strings into which nonempty finite multisets of finite binary strings are encoded.

Theory and applications are given in the textbook [19]. Here we give some relations that are needed in the paper. The *information about* x *contained in* y is defined as $I(y:x) = K(x) - K(x|y)$. A deep, and very useful, result holding for both plain complexity and prefix complexity, due to L.A. Levin and A.N. Kolmogorov [29] called *symmetry of information* states that

$$K(x,y) = K(x) + K(y|x) = K(y) + K(x|y), \tag{II.1}$$

with the equalities holding up to a $O(\log K)$ additive term. Here, $K = \max\{K(x), K(y)\}$. Hence, up to an additive logarithmic term $I(x:y) = I(y:x)$ and we call this the *mutual (algorithmic) information* between x and y.

B. Multiset

A multiset is also known as *bag*, *list*, or *multiple*. A *multiset* is is a generalization of the notion of set. The members are allowed to appear more than once. For example, if $x \neq y$ then $\{x,y\}$ is a set, but $\{x,x,y\}$ and $\{x,x,x,y,y\}$ are multisets. We abuse the common set-membership notation by using it for multisets by writing $x \in \{x,x,y\}$ and $z \notin \{x,x,y\}$ for $z \neq x,y$. Further, $\{x,x,y\} \setminus \{x\} = \{x,y\}$. If X,Y,Z are multisets and X,Z are nonempty and $X = YZ$, then we write $Y \subset X$. For us, a multiset is finite such as $\{x_1,\ldots,x_m\}$ with $m < \infty$ and the members are finite binary strings in length-increasing lexicographic order. If X is a multiset, then some or all of its elements may be equal. Thus, $x_i \in X$ means that "x_i is an element of multiset X." With $\{x_1,\ldots,x_{m+1}\} \setminus \{x\}$ we mean the length-increasing lexicographic concatenation of $x_1 \ldots x_{m+1}$ with one occurrence of x removed.

The finite binary strings, finiteness, and length-increasing lexicographic order allows us to assign a unique Kolmogorov complexity to a multiset. The conditional prefix Kolmogorov complexity $K(X|x)$ of a multiset X given an element x is the length of a shortest program p for the reference universal Turing machine that with input x outputs the multiset X. The prefix Kolmogorov complexity $K(X)$ of a multiset X is defined by $K(X|\epsilon)$. One can also put multisets in the conditional such as $K(x|X)$ or $K(X|Y)$. We will use the straightforward laws $K(\cdot|X,x) = K(\cdot|X)$ and $K(X|x) = K(X'|x)$ up to an additive constant term, for $x \in X$ and X' equals the multiset X with one occurrence of the element x deleted.

C. Information Distance

The information distance in a multiset X ($|X| \geq 2$) is given by (I.1). To obtain the *pairwise information distance* in [2] we take $X = \{x_1, x_2\}$ in (I.1). The resulting formula is equivalent to $E_{\max}(x_1, x_2) = \max\{K(x_1|x_2), K(x_2|x_1)\}$ up to a logarithmic additive term.

D. Metricity

Let $\mathcal{X}$ be the set of length-increasing lexicographic ordered finite multisets of finite binary strings. A *distance function* d on $\mathcal{X}$ is defined by $d : \mathcal{X} \to \mathcal{R}^+$ where $\mathcal{R}^+$ is the set of nonnegative real numbers. Define $Z = XY$ if Z is a multiset of the elements of the multisets X and Y and the elements of Z are ordered length-increasing lexicographic. A distance function d is a *metric* if

1) *Positive definiteness*: $d(X) = 0$ if all elements of X are equal and $d(X) > 0$ otherwise.
2) *Symmetry*: $d(X)$ is invariant under all permutations of X.
3) *Triangle inequality*: $d(XY) \leq d(XZ) + d(ZY)$.

We recall Theorem 4.1 and Claim 4.2 from [24].

Theorem 2.1: The information distance for multisets $E_{\max}$ is a metric where the (in)equalities hold up to a $O(\log K)$ additive term. Here K is the largest quantity involved in each metric (in)equality 1) to 3), respectively.

Claim 2.2: Let X, Y, Z be three multisets of finite binary strings and $K = K(X) + K(Y) + K(Z)$. Then, $E_{\max}(XY) \leq E_{\max}(XZ) + E_{\max}(ZY)$ up to an $O(\log K)$ additive term.

III. Normalized Information Distance

The quantitative difference in a certain feature between many objects can be considered as an *admissible* distance, provided it is upper semicomputable and satisfies a density condition for every $x \in \{0,1\}^*$ (to exclude distances like $D(X) = 1/2$ for every multiset X):

$$\sum_{X: x \in X \ \& \ D(X) > 0} 2^{-D(X)} \leq 1. \tag{III.1}$$

Thus, for the density condition on D we consider only multisets X with $|X| \geq 2$ and not all elements of X are equal. Moreover, we consider only distances that are semicomputable, that is, they are computable in some broad sense (they can be computably approximated from above). Theorem 5.2 in [24] shows that $E_{\max}$ is universal in that among all admissible multiset distances in that it is always least up to

an additive constant. That is, it accounts for the dominant feature in which the elements of the given multiset are alike.

Admissible distances as defined above are absolute, but if we want to express similarity, then we are more interested in relative ones. For example, if a multiset X of strings of each about $1{,}000{,}000$ bits have pairwise information distance $1{,}000$ bits to each other, then we are inclined to think that those strings are relatively similar. But if a multiset Y consists of strings of each about $1{,}200$ bits and each two strings in it have a pairwise information distance of $1{,}000$ bits, then we think the strings in Y are very different. In the first case $E_{\max}(X) \approx 1{,}000|X| + O(1)$, and in the second case $E_{\max}(Y) \approx 1{,}000|Y| + O(1)$. In case $|X| \approx |Y|$ the information distances of the multsets are about the same.

Therefore, to express similarity we need to normalize the universal information distance $E_{\max}$ to obtain a universal similarity distance. It should give a similarity with distance 0 when the objects in a multiset are maximally similar (that is, they are equal) and distance 1 when they are maximally dissimilar. Naturally, we desire the normalized version of the universal multiset information distance metric to be also a metric.

For pairs of objects x, y the normalized version e of $E_{\max}$ defined by

$$e(x,y) = \frac{E_{\max}(x,y)}{\max\{K(x), K(y)\}} = \frac{\max\{K(x,y|x), K(x,y|y\}}{\max\{K(x), K(y)\}} \tag{III.2}$$

takes values in $[0,1]$ up to an additive term of $O(1/K(x,y))$. It is a metric up to additive terms $O((\log K)/K)$ with K denotes the maximum of the Kolmogorov complexities involved in each of the metric (in)equalities, respectively. A normalization factor for multisets of more than two elements ought to reduce to that of (III.2) for multisets restricted to two elements. Consider strings consisting of the concatenation of finite multisets of finite strings.

Remark 3.1: For example set $X = \{x\}, Y = \{y, y\}, Z = \{y\}, K(x) = n, K(x|y) = n, K(y) = 0.9n$ and by using (II.1) we have $K(x,y) = 1.9n, K(y|x) = 0.9n$. The most natural definition is a generalization of (III.2):

$$e(A) = \frac{E_{\max}(A)}{\max_{x \in A}\{K(A \setminus \{x\})\}}.$$

But we find $e(XY) = K(x|y)/K(x,y) = n/1.9n \approx 1/2$, and $e(XZ) = K(x|y)/K(x) = n/n = 1$, $e(ZY) = K(y|y)/K(y) = 0/0.9n = 0$, and the triangle inequality is violated. Intuitively, if we add an element to a multiset of objects then a program to go from any object in the new multiset to any other object should be at least as long as a program to go from any object in the old multiset to any other obkect. This suggests a definition of $e(A)$ that is nondecreasing when we add elements to A.

This leads to (III.3). With $B \subset A$ with $B = A \setminus \{y\}$ we find $e(XY) = K(x|y)/K(x) = n/n = 1$, $e(XZ) = K(x|y)/K(x) = n/n = 1$ and $e(ZY) = 0/n = 0$. The triangle inequality holds. $\diamondsuit$

The reasoning in the remark points the way to go: the definition of $e(X)$ with $|X| \geq 2$ should be monotonic nondecreasing in $|X|$ if we want e to be a metric.

Lemma 3.2: Let U, X be multisets and d be a distance that satisfies the triangle inequality. If $U \subseteq X$ then $d(U) \leq d(X)$.

Proof: Let A, B, C be multisets with $A, B \subseteq C$, and d a distance that satisfies the triangle inequality. Assume that the lemma is false and $d(C) < d(AB)$. Let $D = C \setminus A$. It follows from the triangle inequality that

$$d(AB) \leq d(AD) + d(DB).$$

Since $AD = C$ this implies $d(AB) \leq d(C) + d(DB)$, and therefore $d(C) \geq d(AB)$. But this contradicts the assumption. ■

Definition 3.3: Let X be a multiset. Define the *normalized information distance* (NID) for multiples by $e(X) = 0$ for $|X| = 0, 1$ and

$$e(X) = \max \left\{ \frac{E_{\max}(X)}{\max_{x \in X}\{K(X \setminus \{x\})\}}, \max_{Y \subset X}\{e(Y)\} \right\} \tag{III.3}$$

For $|X| = 2$ the value of $e(X)$ is equivalently given in (III.2),

Thus, (III.3) satisfies the property in Lemma 3.2: If X, Z are multisets and $Z \subseteq X$ then $e(Z) \leq e(X)$. Therefore we can hope to prove the triangle property for (III.3).

Theorem 3.4: For every multiset X we have $0 \leq e(X) \leq 1$ up to an additive term of $O(1/K)$ where $K = K(X)$.

Proof: By induction on $n = |X|$. The theorem is vacuously true for $n = 0, 1$.

Base case: $n = 2$. The definition of $e(X)$ is (III.2). The proof of the lemma for this case is in [16].

Induction $n > 2$: Assume that the lemma is true for the cases $2 \leq |X| < n$. Let $|X| = n$. If $e(X) = \max_{Y \subset X}\{e(Y)\}$ then the lemma holds by the inductive assumption since $|Y| < n$. So assume that

$$e(X) = \frac{\max_{x \in X}\{K(X|x)\}}{\max_{x \in X}\{K(X \setminus \{x\})\}}.$$

Since the numerator is at most the denominator up to an $O(1)$ additive term and the denominator at most $K(X)$ the lemma holds for this case. ■

Remark 3.5: The least value of $e(X)$ is reached if all occurrences of alements of X are equal.

In that case $0 \leq e(X) \leq O(1)/K(X)$. The greatest value $e(X) = 1 + O(1/K(X))$ is reached if $\max_{x\in X}\{K(X|x)\} = \max_{x\in X}\{K(X \setminus \{x\}|x)\} + O(1) = \max_{x\in X}\{K(X \setminus \{x\}) + O(1)$. For example, this happens if the selected conditional, say y, has no consequence in the sense that $K(X \setminus \{y\}|y) = K(X \setminus \{y\}) + O(1)$. This happens if $K(z|y) = K(z)$ for all $z \in X \setminus \{y\}$.

Another matter is the consequences of Theorem 3.4. Using (II.1) in the first equality in both the numerator and the denominator. Then we obtain up to additive constants in the numerator and denominator

$$\begin{aligned}\frac{\max_{x\in X}\{K(X|x)\}}{\max_{x\in X}\{K(X \setminus \{x\})\}} &= \frac{K(X) - \min_{x\in X}\{K(x)\}}{K(X) - \min_{x\in X}\{K(x|X \setminus \{x\})\}} \\ &= 1 - \frac{\min_{x\in X}\{K(x)\} - \min_{x\in X}\{K(x|X \setminus \{x\})\}}{K(X) - \min_{x\in X}\{K(x|X \setminus \{x\})\}}.\end{aligned}$$

This expression goes to 1 if both

$$K(X) \to \infty, \ \frac{\min_{x\in X}\{K(x)\}}{K(X)} \to 0.$$

This happens, for instance, if $X \to \{0,1\}^n$ and $n \to \infty$. Another example is $|X| = n$, $\min_{x\in X} = 0$, $K(X) > n^2$, and $n \to \infty$. One can only have $K(X) \to \infty$ and $\min_{x\in X}\{K(x)\}/K(X) \to 0$ if $\min_{x\in X}\{K(x)\} = o(K(X))$ and $\max_{x\in X}\{K(x)\} = \Omega(K(X))$, that is, if X consists of at least two elements and gap between the minimum Kolmogorov complexity and the maximum Kolmogorov complexity of the elements grows to infinity when $K(X) \to \infty$. $\diamondsuit$

Definition 3.6: Let x_u and x_V be defined such that $K(U \setminus \{x_u\}) = \max_{x\in U}\{K(U \setminus \{x\})\}$, and $K(V|x_V) = \max_{x\in V}\{K(V|x)\}$.

Theorem 3.7: For X is a multiset. The function $e(X)$ is a metric up to an additive $O((\log K)/K)$ term in the respective metric (in)equalities, where K is the largest Kolmogorov complexity involved the (in)equality.

Proof: The quantity $e(X)$ satisfies positive definiteness and symmetry up to an $O((\log K(X))/K(X))$ additive term, as follows directly from the definition of $e(X)$. It remains to prove the triangle inequality:

Let X, Y, Z be finite multisets. Then, $e(XY) \leq e(XZ) + e(ZY)$ within an additive term of $O((\log K)/K)$ where $K = \max\{K(X), K(Y), K(Z)\}$.

The proof proceeds by induction on $n = |XY|$. The cases $n = 0, 1$ are vacuously true.

Base case $n = 2$. The definition of $e(XY)$ with XY is a multiset of cardinality 2 is (III.2). The proof of the lemma for this case is in [16].

Induction $n > 2$. Assume that the lemma is true for the cases $2 \leq |XY| < n$. Let $|XY| = n$. If $e(XY) = \max_{Z \subset XY}\{e(Z)\}$ then the lemma holds by the inductive assumption since $|Z| < n$. So assume that

$$e(XY) = \frac{K(XY|x_{XY})}{K(XY \setminus \{x_{xy}\})}.$$

Claim 3.8: Let X, Y, Z be nonempty multisets. $K(XYZ|x_{XYZ}) \leq K(XZ|x_{XZ}) + K(ZY|x_{ZY})$ up to an additive $O(\log K)$ term, where $K = K(X) + K(Y) + K(Z)$.

Proof: By Theorem 2.1 we have that $E_{\max}$ is a metric. In particular, the triangle inequality is satisfied by Claim 2.2: $K(XY|x_{XY}) \leq K(XZ|x_{XZ}) + K(ZY|x_{ZY})$ for multisets X, Y, Z up to an additive term of $O(\log K)$ where $K = K(X) + K(Y) + K(Z)$. Thus with $X' = XZ$ and $Y' = ZY$ we have $K(X'Y'|x_{X'Y'}) \leq K(X'Z|x_{X'Z}) + K(ZY'|x_{ZY'})$ up to the logarithmic additive term. Writing this out $K(XZZY|x_{XZZY}) \leq K(XZZ|x_{XZZ}) + K(ZYZ|x_{ZYZ})$ or $K(XYZ|x_{XYZ}) = K(XZZY|x_{XZZY}) \leq K(XZ|x_{XZ}) + K(ZY|x_{ZY})$ up to an additive term of $O(\log K)$. ■

Now consider the following sequence of inequalities:

$$\begin{aligned} \frac{K(XYZ|x_{XYZ})}{K(XYZ \setminus \{x_{xyz}\})} &\leq \frac{K(XZ|x_{XZ})}{K(XYZ \setminus \{x_{xyz}\})} + \frac{K(ZY|x_{ZY})}{K(XYZ \setminus \{x_{xyz}\})} \\ &\leq \frac{K(XZ|x_{XZ})}{K(XZ \setminus \{x_{xz}\})} + \frac{K(ZY|x_{ZY})}{K(ZY \setminus \{x_{zy}\})}, \end{aligned} \tag{III.4}$$

up to a $O((\log K)/K)$ additive term. The first inequality is Claim 3.8 (by this inequality the denominator is unchanged); the second inequality follows from $K(XYZ \setminus \{x_{xyz}\}) \geq K(XZ \setminus \{x_{xz}\})$ and $K(XYZ \setminus \{x_{xyz}\}) \geq K(ZY \setminus \{x_{zy}\})$ using the principle that $K(u, v) \geq K(u) + O(1)$, reducing both denominators increases the sum of the quotients (by this inequality the numerators are unchanged).

Claim 3.9: Let A be a multiset. If $e(A) < 1$ then there is an B with $A \subset B$ such that $e(A) < e(B)$.

Proof: Let $\epsilon = 1 - e(A)$. Define $B = A \bigcup \{b\}$ such that $e(B) \geq 1 - \epsilon/2$. Namely, with $K(b) > |b|$ and the length of b growing, we have by (III.3) that $e(B) > e(A)$. In particular, we can choose b such that $e(B) \geq e(A) + \epsilon/2$. ■

Assume $e(XY) = 1$. Then $e(XYZ) = 1$ since $e(A) \leq 1$ for every multiset A and by (III.3) we have $e(B) \geq e(A)$ for $A \subset B$. Assume $e(XY) < 1$. By Claim 3.9 we have $e(XY) < e(XYU)$ for some nonempty multiset U. Hence by the definition (III.3) there is a nonempty multiset Z with $Z \subseteq U$ such

that

$$\frac{K(XY|x_{XY})}{K(XY \setminus \{x_{xy}\})} \leq \frac{K(XYZ|x_{XYZ})}{K(XYZ \setminus \{x_{xyz}\})}$$

Together with (III.4) this proves the triangle inequality up to an additive term of $O((\log K)/K)$. ■

By Theorems 3.4 and 3.7 the distance according to (III.3) is a metric with values in $[0,1]$ up to some ignorable additive terms..

IV. Compression Distance for Multisets

We develop the compression-based equivalence of the Kolmogorov complexity based theory in the preceding sections. This is similar to [4] for the case $|X| = 2$. We assume the notion of the real-world compressor G used in the sequel is "normal" in the sense of [4]. By $G(x)$ we mean the length of string x when compressed by G. Consider a multiset X as a string consisting of the concatenated strings of its members ordered length-increasing lexicographic. Thus we can write $G(X)$.

Let $X = \{x_1, \ldots, x_m\}$. The information distance $E_{\max}(X)$ can be rewritten as

$$\max\{K(X) - K(x_1), \ldots, K(X) - K(x_m)\}, \tag{IV.1}$$

within logarithmic additive precision, by (II.1). The term $K(X)$ represents the length of the shortest program for X. The order of the members of X makes only a small difference; block-coding based compressors are symmetric almost by definition, and experiments with various stream-based compressors (gzip, PPMZ) show only small deviations from symmetry.

Approximation of $E_{\max}(X)$ by a compressor G is straightforward: it is

$$E_{G,\max}(X) = \max\{G(X) - G(x_1), \ldots, G(X) - G(x_m)\} = G(X) - \min_{x \in X}\{G(x)\}. \tag{IV.2}$$

We need to show it is an admissible distance and a metric.

Lemma 4.1: If G is a normal compressor, then $E_{G,\max}(X)$ is an admissible distance.

Proof: For $E_{G,\max}(X)$ to be an admissible distance it must satisfy the density requirement (III.1) and be upper semicomputable. Since the length $G(x)$ is computable it is a fortiori upper semicomputable. The density requirement (III.1) is equivalent to the Kraft inequality [14] and states in fact that for every string x the set of $E_{G,\max}(X)$ is a prefix-free code for the X's containing x. According to (IV.2) we have for any fixed $x \in X$: $E_{G,\max}(X) \geq G(X) - G(x) \geq G(X \setminus \{x\})$. Hence, $2^{-E_{G,\max}(X)} \leq 2^{-G(X \setminus \{x\})}$

and therefore

$$\sum_{X:x\in X} 2^{-E_{G,\max}(X)} \leq \sum_{X:x\in X} 2^{-G(X\setminus\{x\})}.$$

A compressor G compresses strings into a uniquely decodable code (it must satisfy the unique decompression property) and therefore the length set of the compressed strings must satisfy the Kraft inequality [20]. Then, for a fixed given x the compressed code for the multisets $X \setminus \{x\}$ must satisfy this inequality. Hence the right-hand side of above displayed inequality is at most 1. ■

Lemma 4.2: If G is a normal compressor, then $E_{G,\max}(X))$ is a metric with the metric (in)equalities satisfied up to logarithmic additive precision.

Proof: Let X, Y, Z be multisets with at most m members of length at most n. The positive definiteness and the symmetry property hold clearly up to an $O(\log G(X)))$ additive term. Only the triangular inequality is nonobvious. For every compressor G we have $G(XY) \leq G(X) + G(Y)$ up to an additive $O(\log(G(X) + G(Y)))$ term, otherwise we obtain a better compression by dividing the string to be compressed. (This also follows from the distributivity property of normal compressors.) By the monotonicity property $G(X) \leq G(XZ)$ and $G(Y) \leq G(YZ)$ up to an $O(\log(G(X) + G(Y)))$ or $O(\log(G(Y) + G(Z)))$ additive term, respectively. Therefore, $G(XY) \leq G(XZ) + G(ZY)$ up to an $O(\log(G(X) + G(Y) + G(Z)))$ additive term. ■

V. Normalized Compression Distance for Multisets

Let X be a multiset. The normalized version of $e(X)$ using the compressor G based approximation of the normalized information distance for multisets (III.3), is called the *normalized compression distance* (NCD) for multisetsmultiples: $NCD(X) = 0$ for $|X| = 0, 1$; if $|X| \geq 2$ then

$$NCD(X) = \max\left\{\frac{G(X) - \min_{x\in X}\{G(x)\}}{\max_{x\in X}\{G(X\setminus\{x\}\}}, \max_{Y\subset X}\{NCD(Y)\}\right\}. \tag{V.1}$$

Here $X \setminus \{x\}$ denotes the string consisting of the length-increasing lexicographic elements of X with one occurrence of the substring x removed.

This NCD is the main concept of this work. It is the real-world version of the ideal notion of normalized information distance NID for multiples in (III.3).

Remark 5.1: In practice, the NCD is a non-negative number $0 \leq r \leq 1 + \epsilon$ representing how different the two files are. Smaller numbers represent more similar files. The ϵ in the upper bound is due to imperfections in our compression techniques, but for most standard compression algorithms one is unlikely

to see an ϵ above 0.1 (in our experiments gzip and bzip2 achieved NCD's above 1, but PPMZ always had NCD at most 1). $\diamondsuit$

Theorem 5.2: If the compressor is normal, then the NCD for multiples is a normalized admissible distance and satisfies the metric (in)equalities up to an ignorable additive term, that is, a similarity metric.

Proof: The NCD (V.1) is a normalized admissible distance by Lemma 4.1. It is normalized to $[0,1]$ up to an additive term of $O(1/G)$ with $G = G(X)$ as we can see from the formula (V.1) and Theorem 3.4 with G substituted for K throughout. We next show it is a metric.

A normal compressor is idempotent in the sense that $NCD(X) = 0$ if X consists of equal members. The idempotency property of a normal compressor is up to an additive term of $O(\log G(X))$. Hence the positive definiteness of $G(X)$ is satisfied up to an additive term of $O((\log G(X))/G(X))$. The order of the members of X is assumed to be length-increasing lexicographic. Therefore it is symmetric up to an additive term of $O((\log G(X))/G(X))$. It remains to show the triangle inequality $NCD(XY) \leq NCD(XZ) + NCD(ZY)$ up to an additive term of $O((\log G)/G)$ where $G = G(X) + G(Y) + G(Z)$. We do this by induction on $n = |XY|$ where X, Y are multisets. For $n = 0, 1$ the triangle property is vacuously satisfied.

Base case $n = 2$. That is, $|XY| = 2$. This is proved in [4].

Induction $n > 2$. Assume the triangle property is satisfied for $2 \leq |XY| < n$. Then we prove it for $|XY| = n$. If $NCD(XY) = NCD(Z)$ for some $Z \subset XY$ then $2 \leq |Z| < n$ and the case follows from the inductive argument. Therefore, $NCD(XY)$ is the first term in the outer maximization of (V.1). Write $G(XY|x_{XY}) = G(XY) - \min_{x \in XY}\{G(x)\}$ and $G(XY \setminus \{x_{xy}\}) = \max_{x \in XY}\{G(XY) \setminus \{x\}\}$ and similar for XZ, YZ, XYZ. Following the proof of the induction case of the triangle inequality in the proof of Theorem 3.7, using Lemma 4.2 for the metricity of $E_{G,\max}$ wherever Theorem 2.1 is used to assert the metricity of $E_{\max}$, and substitute G for K in the remainder. This completes the proof. That is, for every multiset Z we have

$$NCD(XY) \leq NCD(XZ) + NCD(ZY),$$

up to an additive term of $O((\log G)/G)$. ■

VI. Computing the Normalized Compression Distance for Multsets

In practice it seems that one can do no better than follow the definition inductively. Assume we want to compute $NCD(X)$ and $|X| \geq 2$.

Base Step: Compute $M_2 = \max_{Y \subset X,\, |Y|=2}\{NCD(Y)\}$, as in [4].

Induction: $2 \leq m < n$. Let $M_m = \max_Z\{e(Z) : Z \subset X,\, 2 \leq |Z| \leq m\})$. Then with $Z \subset Y \subseteq X$ and $|Y| = m + 1$:

$$NCD(Y) = \max\left\{\frac{G(Y) - \min_{x \in Y}\{G(x)\}}{\max_{x \in Y}\{G(Y \setminus \{x\})\}}, M_m\right\}.$$

We ignore logarithmic additive terms. With $|Y| = n$ we have $Y = X$ and hence $NCD(Y) = NCD(X)$. However, this process involves evaluating the NCD's of the almost the entire powerset of X.

Natural Data and Kolmogorov Complexity: The Kolmogorov complexity of a file is a lower bound on the length of the ultimate compressed version of that file. In both cases above we approximate the Kolmogorov complexities involved by a real-world compressor. Since the Kolmogorov complexity is incomputable, in the approximation we never know how close we are to it. However, we assume that the natural data we are dealing with contain no complicated mathematical constructs like $\pi = 3.1415\ldots$ or Universal Turing machines. In fact, we assume that the natural data we are dealing with contains only effective regularities that a good compressor finds. Under those assumptions the Kolmogorov complexity of the object is not much smaller than the length of the compressed version of the object.

VII. Applications

We detail preliminary results using the new NCD for multiples. The NCD for pairs as originally defined [4] has been applied in a wide range of application domains. In [9] a close relative was compared to every time series distance measure published in the decade preceeding 2004 from all of the major data analysis conferencea and found to outperform all other distances aside from the Euclidean distance with which it was competitive. The NCD for pairs has also been applied in biological applications to analyze the results of segmentation and tracking of proliferating cells and organelles [6], [7], [26]. Here, we compare the performance of the proposed NCD for multiples to that of a previous application of the NCD for pairs for predicting retinal progenitor cell (RPC) fate outcomes from the segmentation and tracking results from live cell imaging. We also apply the proposed NCD to a synthetic time sequence data set [10].

A. Retinal Progenitor Cell Fate Prediction

In [7], long-term time-lapse image sequences showing rat RPCs were analyzed using automated segmentation and tracking algorithms. Images were captured every five minutes of the RPCs for a period of 9–13 days. Up to 100 image sequences may be captured simultaneously in this manner using a microscope with a mechanized stage. At the conclusion of the experiment, the "fate" of the offspring produced by each RPC was determined using a combination of cell morphology and specific cell-type fluorescent markers for the four different retinal cell types produced from embryonic day 20 rat RPCs [3]. At the conclusion of the imaging, automated segmentation and tracking algorithms [25] were applied to extract the time course of features for each cell. These automated segmentation and tracking algorithms extract a time course of feature data for each stem cell at a five-minute temporal resolution, showing the patterns of cellular motion and morphology over the lifetime of the cell. Specifically, the segmentation and tracking results consisted of a 6-dimensional time sequence feature vector incorporating two-dimensional motion $(d\ x, d\ y)$, as well as the direction of motion, total distance travelled, cellular size or area (in pixels) and a measure of eccentricity on $[0, 1]$ (0 being linear, 1 being circular shape). The time sequence feature vectors for each of the cells are of different length and are not aligned. The results from the segmentation and tracking algorithms were then analyzed as follows.

The original analysis of the RPC segmentation and tracking results used a multiresolution semi-supervised spectral analysis based on the originally formulated pairwise NCD. An ensemble of distance matrices consisting of pairwise NCDs between quantized time sequence feature vectors of individual cells is generated for different feature subsets f and different numbers of quantization symbols n for the numerical time sequence data. The fully automatic quantization of the numeric time sequence data is described in [6]. All subsets of the 6-dimensional feature vector were included, although it is possible to use non-exhaustive feature subset selection methods such as forward floating search, as described in [6]. Each distance matrix is then normalized as described in [7], and the eigenvectors and eigenvalues of the normalized matrix are computed. These eigenvectors are stacked and ordered by the magnitude of the corresponding eigenvalues to form the columns of a new "spectral" matrix. The spectral matrix is a square matrix, of the same dimension N as the number of stem cells being analyzed. The spectral matrix has the important property that the ith row of the matrix is a point in $\mathcal{R}^N$ ($\mathcal{R}$ is the set of real numbers) that corresponds to the quantized feature vectors for the ith stem cell. If we consider only the first k columns, giving a spectral matrix of dimension $N \times k$, and run a K-Means clustering algorithm, this yields the well-known spectral K-Means algorithm [8]. If we have known outcomes for any of the

objects that were compared using the pairwise NCD, then we can formulate a semi-supervised spectral learning algorithm by running for example nearest neighbors or decision tree classifiers on the rows of the spectral matrix. This was the approach adopted in [7].

In the original analysis, three different sets of known outcomes were considered. First, a group of 72 cells were analyzed to identify cells that would self-renew (19 cells), producing additional progenitors and cells that would terminally differentiate (53 cells), producing two retinal neurons. Next, a group of 86 cells were considered on the question of whether they would produce two photoreceptor neurons after division (52 cells), or whether they would produce some other combination of retinal neurons (34 cells). Finally, 78 cells were analyzed to determine the specific combination of retinal neurons they would produce, including 52 cells that produce two photoreceptor neurons, 10 cells that produce a photoreceptor and bipolar neuron, and 16 cells that produced a photoreceptor neuron and an amacrine cell. For the terminal versus self-renewing question, 99% accuracy was achieved in prediction using a spectral nearest neighbor classifier. For the two photoreceptor versus other combination question, 87% accuracy was achieved using a spectral decision tree classifier. Finally, for the specific combination of retinal neurons 83% accuracy was achieved also using a spectral decision tree classifier.

Classification using the newly proposed NCD (III.3) is much more straightforward and leads to significantly better results. Given multisets A and B, each consisting of cells having a given fate, and a cell x with unknown fate, we proceed as follows. We assign x to whichever multiset has its distance (more picturesque "diameter") increased the least with the addition of x. In other words, if

$$NCD(Ax) - NCD(A) < NCD(Bx) - NCD(B), \tag{VII.1}$$

we assign x to multiset A, else we assign x to multiset B. (The notation Xx is shorthand for the multiset X with one occurrence of x added.) For this first dataset, we did not need to evaluate the subset term, or second term in the outer maximization in (III.3) as detailed in the following section.

The classification accuracy improved considerably using the newly proposed NCD for multiples. For the terminal versus self-renewing question, we achieved 100% accuracy in prediction compared to 99% accuracy for the multiresolution spectral pairwise NCD. For the two photoreceptor versus other combination question, we also achieved 100% accuracy compared to 87%. Finally, for the specific combination of retinal neurons we achieved 92% accuracy compared to 83% with the previous method.

B. Synthetic Time Sequence Data

In addition to the retinal progenitor cell data, we applied the new NCD for multiples to analyzing synthetically generated time sequence data intended for the characterization of new machine learning algorithms [10]. The testing data here consists of 300 numerical time sequences, each containing 60 time points and belonging to one of six classes. This data is classified with 88% accuracy using a nearest neighbor Euclidian distance classifier. (We note that other methods such as dynamic time warping (DTW) have achieved considerably better results.)

In applying the NCD to this data, we first measure the separation between classes or the *margin*. Given multisets A and B, each corresponding to a class in the testing data, we measure the separation between the two classes as

$$NCD(AB) - NCD(A) - NCD(B). \tag{VII.2}$$

This follows directly from the relevant Venn diagram. Our approach is to ensure that the separation between classes is larger than any separation between subsets of the same class. If the separation between classes is larger than the separation within any class, we can ignore the subset component of (III.3). Verifying this requires us to evaluate the NCD over the powerset of each class and is this is not feasible. Instead, we have developed an approximate approach based on an expectation maximization algorithm to partition the classes such that there exist no subsets of a class separated by a margin larger than the minimum separation between classes.

Our expectation maximization algorithm attempts to partition the classes into maximally separated subsets as measured by (VII.2). This algorithm, that we have termed *K-Lists*, is modeled after the K-means algorithm. Although it is suitable for general clustering, here we use it to partition the data into two maximally separated subsets. The algorithm is detailed in Figure 1. There is one important difference between proposed K-Lists algorithm and the K-Means algorithm. Because we are not using the centroid of a cluster as a representative value as in K-Means, but rather the subset itself via the NCD for multiples, we only allow a single element to change subsets at every iteration. This prevents thrashing where groups of elements chase each other back and forth between the two subsets. This step is computationally demanding, but it is an inherently parallel computation.

For the retinal progenitor cell data described in the previous section, the K-Lists partitioning algorithm was not able to find any subsets for any of the three questions that had a larger separation as measured by (VII.2) compared to the separation between the classes. For the synthetic data, the partitioning algorithm

1) (Initialize) Pick two elements (seeds) of X at random, assigning one element to each A and B. For each remaining element x, assign x to the closer one of A or B using pairwise NCD to the random seeds
2) For each element x, compute the distance from x to class A and B using (VII.1) and assign to whichever class achieves the smaller distance.
3) Choose the single element that wants to change subsets, e.g. from A to B or vice versa and whose change maximizes $NCD(AB) - NCD(A) - NCD(B)$ and swap that element from A to B or vice versa.
4) Repeat steps 2 and 3 until no more elements want to change subsets or until we exceed e.g. 100 iterations.

Repeat the whole process some fixed number of times (here we use 5) for each X and choose the subsets that achieve the maximum of $e(AB) - e(A) - e(B)$. If that value exceeds the minimum inter-class separation then divide X into A and B and repeat the process for A and B. If the value does not exceed the minimum inter-class separation of our training data, then accept X as approximately monotonic and go on to the next class.

Fig. 1. Partitioning algorithm for identifying maximally separated subsets For each class (multiset) X, partition X into two subsets A and B such that $NCD(AB) - NCD(A) - NCD(B)$ is a maximum

was consistently able to find subsets with separation larger than the between class separation. For the synthetic data, the partitioning was run repeatedly and the best partitioning selected using cross validation. The accuracy of this preliminary classification was 88% correct. This is equivalent to the nearest neighbor Euclidean-distance classifier, and less accurate than the classifiers that used DTW.

C. Data, Software, Machines

All of the software and the time sequence data for the RPC fate outcome problem can be downloaded from http://bioimage.coe.drexel.edu/ncdm. The data for the synthetic time sequence can be downloaded from [10]. The software is implemented in C and uses MPI for parallelization. Data import is handled by a MATLAB script that is also provided. The software has been run on a very small cluster, consisting of 150 (hyperthreaded) Xeon and I7 cores running at 2.9 Ghz. The RPC classification runs in approximately 20 minutes for each question, while the partitioning and classification of the synthetic data takes a few hours.

REFERENCES

[1] C. Ané and M. Sanderson, Missing the forest for the trees: phylogenetic compression and its implications for inferring complex evolutionary histories, *Systematic Biology*, 54:1(2005), 146–157.

[2] C.H. Bennett, P. Gács, M. Li, P.M.B. Vitányi, and W. Zurek. Information distance, *IEEE Trans. Inform. Theory*, 44:4(1998), 1407–1423.

[3] M. Cayouette, B. A. Barres, and M. Raff, Importance of intrinsic mechanisms in cell fate decisions in the developing rat retina, *Neuron*, 40(2003), 897–904.

[4] R.L. Cilibrasi, P.M.B. Vitányi, Clustering by compression, *IEEE Trans. Inform. Theory*, 51:4(2005), 1523- 1545.

[5] R.L. Cilibrasi, P.M.B. Vitányi, The Google similarity distance, *IEEE Trans. Knowledge and Data Engineering*, 19:3(2007), 370-383.

[6] A. R. Cohen, C. Bjornsson, S. Temple, G. Banker, and B. Roysam, Automatic summarization of changes in biological image sequences using Algorithmic Information Theory, *IEEE Trans. Pattern Anal. Mach. Intell.* 31(2009), 1386–1403.

[7] A. R. Cohen, F. Gomes, B. Roysam, and M. Cayouette, "Computational prediction of neural progenitor cell fates, *Nature Methods*, 7(2010), 213–218.

[8] S. D. Kamvar, D. Klein, and C. D. Manning, Spectral learning, Proc. Int. Joint Conf. Artificial Intelligence, 2003, 561–566.

[9] E. Keogh, S. Lonardi, and C. A. Ratanamahatana, Towards parameter-free data mining, Proc. 10th ACM SIGKDD Int. Conf. Knowledge Discovery and Data Mining, 2004,206–215.

[10] E. Keogh, Q. Zhu, B. Hu, Y. Hao, X. Xi, L. Wei and C.A. Ratanamahatana, The UCR Time Series Classification/Clustering Homepage: www.cs.ucr.edu/∼eamonn/time_series_data/, 2011.

[11] S.R. Kirk and S. Jenkins, Information theory-based software metrics and obfuscation, *Journal of Systems and Software*, 72(2004), 179-186.

[12] A. Kocsor, A. Kertész-Farkas, L. Kaján, and S. Pongor, Application of compression-based distance measures to protein sequence classification: a methodology study, *Bioinformatics*, 22:4(2006), 407–412.

[13] A.N. Kolmogorov, Three approaches to the quantitative definition of information, *Problems Inform. Transmission* 1:1(1965), 1–7.

[14] L.G. Kraft, A device for quantizing, grouping, and coding amplitude modulated pulses, MS Thesis, EE Dept., Massachusetts Institute of Technology, Cambridge. Mass., USA.

[15] L.A. Levin, Laws of information conservation (nongrowth) and aspects of the foundation of probability theory, *Probl. Inform. Transm.*, 10(1974), 206–210.

[16] M. Li, X. Chen, X. Li, B. Ma, P.M.B. Vitányi. The similarity metric, *IEEE Trans. Inform. Theory*, 50:12(2004), 3250-3264.

[17] M. Li, C. Long, B. Ma, X. Zhu, Information shared by many objects, Proc. 17th ACM Conf. Information and Knowledge Management, 2008, 1213–1220.

[18] M. Li, Information distance and its extensions, Proc. Discovery Science, Lecture Notes in Computer Science, Volume 6926, 2011, 18–28

[19] M. Li and P.M.B. Vitányi. *An Introduction to Kolmogorov Complexity and its Applications*, Springer-Verlag, New York, Third edition, 2008.

[20] B. McMillan, Two inequalities implied by unique decipherability, *IEEE Trans. Information Theory*, 2:4(1956), 115-116.

[21] An.A. Muchnik, Conditional complexity and codes, *Theor. Comput. Sci.*, 271(2002), 97–109.

[22] M. Nykter, N.D. Price, M. Aldana, S.A. Ramsey, S.A. Kauffman, L.E. Hood, O. Yli-Harja, and I. Shmulevich, Gene expression dynamics in the macrophage exhibit criticality, *Proc. Nat. Acad. Sci. USA*, 105:6(2008), 1897–1900.

[23] M. Nykter, N.D. Price, A. Larjo, T. Aho, S.A. Kauffman, O. Yli-Harja and I. Shmulevich, Critical networks exhibit maximal information diversity in structure-dynamics relationships, *Physical Review Lett.*, 100(2008), 058702(4).

[24] P.M.B. Vitányi, Information distance in multiples, *IEEE Trans. Inform. Theory*, 57:4(2011), 2451-2456.

[25] M. Winter, E. Wait, B. Roysam, S. Goderie, E. Kokovay, S. Temple, et al., Vertebrate neural stem cell segmentation, tracking and lineaging with validation and editing, *Nature Protocols*, 6(2011), 1942–1952.

[26] M. R. Winter, C. Fang, G. Banker, B. Roysam, and A. R. Cohen, Axonal transport analysis using Multitemporal Association Tracking, *Int. J. Comput. Biol. Drug Des.*, 5(2012), 35–48.

[27] W. Wong, W. Liu, M. Bennamoun, Featureless Data Clustering, pp 141–164 (Chapter IX) in: *Handbook of Research on Text and Web Mining Technologies*, Idea Group Inc., 2008.

[28] X. Zhang, Y. Hao, X. Zhu, M Li, Information distance from a question to an answer, Proc. 13th ACM SIGKDD Int. Conf. Knowledge Discovery and Data Mining, 2007, 874–883.

[29] A.K. Zvonkin and L.A. Levin, The complexity of finite objects and the development of the concepts of information and randomness by means of the theory of algorithms, *Russian Math. Surveys* 25:6 (1970) 83-124.