圧縮を用いた類似度判定のための計算実験

Experiments for compression based similarity distance

谷　研究室　　新井　秀
Shu Arai

概要

圧縮に基づいたデータ間の類似度に関する距離が定義され，ＤＮＡ、言語，音楽の類似度判定に有用だという実験結果が得られている．本研究では圧縮に基づいた類似度判定についての計算実験を行う。

1　はじめに

あらゆる事柄は、常に何かに分類されることによって人々に認識される。今日、コンピュータの急速な発達により音声や文章はデジタルなデータとして保存されている。

Ming Li，Xin Chen，Bin Ma，Paul M.B. Vitan yi らは背景知識を必要としない万能なデータ間の類似度距離として Kolmogorov 記述量を用いる NID(Normalized Information Distance) を提案した。しかし、Kolmogorov 記述量は計算不可能である。そこで、Kolmogorov 記述量の代用として圧縮アルゴリズムを用いた NCD(normalized compression distance) が提案された。NCD を用いると DNA の類似度や言語の類似度、音楽の類似度判定に有用だということが分かっている。

本研究では NCD を用いた類似度判定についての計算実験を行い、類似度判定における適当な圧縮アルゴリズムやデータについて研究していく。

第 2 節では Kolmogorov 記述量について，第 3 節では NID およびＮＣＤについて述べ，第４節では、様々な実験を行い、結果を考察していく。

2　Kolmogorov 記述量の定義について

この節では Kolmogorov 記述量の基本的な定義について、述べる。

2.1　Kolmogorov 記述量

定義

データ x の Kolmogorov 記述量は、直感的には、あるプログラム言語で x を生成する最小のプログラムサイズのことである。ただし，この場合プログラムとは，空の状態からスタートし，データをすべてを出力し停止するものとする．どんなプログラミング言語を選んでも，それが妥当であれば，情報量は定数分の差しかないことが証明されている．

プログラム言語 S による、データ (x) の kolmogorov 記述量 $K_s(x)$ は以下と定義される。

$$K_s(x) = min\{|p| : S(p) = x\}$$

次に補助情報 y に対するデータ x の Kolomogorov 記述量 $k(x|y)$ を U はデータを記述する言語として固定した万能言語であるとして以下と定義する。

$$K(x|y) = min\{|p| : U(p, y) = x\}$$

2.2　Kolmogorov 記述量のアルゴリズム的性質

$K(x)$ を正の整数から正の整数への関数と考える。

定理

(a) 関数 $K(x)$ は帰納的ではない。しかも、どのような機能的部分関数 $\phi(x)$ を考えても、もしその値が無限個の点で定義されているのならば、その定義上のどこかの点で $\phi(x) \neq K(x)$ となる。

(b) 引数 t に対しては単調減少で（全域的な）機能的関数 $H(t, x)$ が存在し,$\lim_{t\to\infty} H(t, x) = K(x)$。すなわち、$K(x)$ の良い近似は計算可能。ただし一様近似ではない。

2.3　情報量

補助情報 y に生成したい文字列 x に関する情報が多く含まれているとすると $K(x|y)$ は小さくなると思われる。x を生成する最小のプログラムを x^* で表す。つまり、$K(x) = |x^*|$ となる。

x に含まれる y に関する情報量 (inforamtion in x about y) I(x:y) を以下と定義する。

$$I(x : y) = K(x) - K(x|y^*)$$

明らかに $K(x)$=$K(x|y^*)$ であるが、$K(x)$ から $K(x|y^*)$

に減った分を y に含まれる x に関する情報（量）としようというのである。I（x:y）と I(y:x) は完全に一致するとは限らないが、ある定数 c が存在して、I（x:y）-I（y:x）＜ c となることが知られている。つまり、定数の差を無視すれば Kolmogorov 記述量に基づき定義した情報 I は対称性を持つ（I（x:y）＝I（y:x）である）と見なせるのである。すなわち、定数の差を無視すれば

$$K(x) + K(y|x^*) = K(y) + K(x|y^*)$$

3 ＮＩＤおよびＮＣＤについて

3.1 距離

数学において距離空間とは，任意の 2 点間で距離が定められた空間のことをいう.

定義

ある集合 X 上の距離とは，実数値関数 $d : X \times X \to R$ で任意の $x, y, z \in X$ に対して次のような性質を満たす。

$$d(x, y) \geq 0$$
$$d(x, y) = 0 \Leftrightarrow x = y$$
$$d(x, y) = d(y, x) : \text{対称性}$$
$$d(x, y) \leq d(x, z) + d(z, y) : \text{三角不等式}$$

3.2 NID(Normalized Information Distance)

Ming Li らの研究では，類似度に関する距離を正規化し、その距離をＮＩＤ (Noramalized Inforamation Distance) と呼ぶ。

定義

任意のデータ x, y について，以下のようにＮＩＤを定義する。

$$NID(x, y) = \frac{max\{K(x|y^*), K(y|x^*)\}}{max\{K(x), K(y)\}}$$

$K(x) \leq K(y)$ として、$I(x : y)$ を用いると

$$NID(x, y) = \frac{K(y) - I(x : y)}{K(y)} = 1 - \frac{I(x : y)}{K(y)}$$

3.3 NCD(Normalized Compression Distance)

Kolmogorov 記述量は計算不可能なため実際に NID を求めることは出来ない。K の近似として圧縮アルゴリズムを用いた類似度距離 NCD を定義する。

定義

x を文字列として、C(x) をある圧縮アルゴリズムで圧縮したサイズとする。また、xy を x と y の連接とする。

$$NCD(x, y) = \frac{C(xy) - min\{C(x), C(y)\}}{max\{C(x), C(y)\}}$$

また，$C(y) \geq C(x)$ としたとき，

$$NCD(x, y) \quad = \quad \frac{C(xy) - C(x)}{C(y)}$$

NCD は 0 以上、１以下の値をとる。０に近ければ近いほど２つのデータは類似度が高いと判定され、１に近づくとデータは類似度が低いと判定される。

3.4 CDM(compression-based dissimilarity measure)

NCD の簡略版として、CDM(compression-based dissimilarity measure) という測定方法も提案されている。

定義

$$CDM(x, y) = \frac{C(xy)}{C(x) + C(y)}$$

ただし、$\frac{1}{2}$のとき、類似度が高く、1 のとき低い。 (1)

4 実験

4.1 ncd と cdm で計算される類似度の違い

ncd の簡略版として cdm が提案された。しかし、その精度については個人的にも疑わしいものがある。同じファイルの類似度を両方で調べることによって、cdm の精度を調べる。

実験方法

使用するファイル：方言桃太郎の前処理後のファイル
圧縮アルゴリズム:gzip
比較方法：ncd と cdm 両方の類似度を調べ。その差で誤差の割合 (%) で出す。
結果は以下の画像 (一部)。

0	1.071754	0.088101	2.135584	2.038369	0.735457	0.518135
1.13321	0	1.498605	1.160621	1.08293	2.235962	1.830233
0.088101	1.536361	0	2.032466	2.234132	0.706861	0.442462
2.077268	1.19277	2.032466	0	0.260756	2.910745	2.426433
1.984127	1.05521	2.176209	0.260756	0	3.473941	3.36497
0.735457	2.291326	0.706861	2.910745	3.556663	0	0.368411
0.490315	1.783778	0.430663	2.564103	3.524799	0.351824	0
1.718295	0.688509	1.849949	0.630135	0.580287	2.805543	2.74988
1.634277	0.766551	1.958103	0.509804	0.373155	2.921301	2.355072
0.164957	1.480611	0.084631	2.284382	2.762431	0.736428	0.488334
1.05251	2.624672	1.076142	3.333333	4.55564	0.458141	1.006392
0.598855	2.326914	0.569158	2.849315	4.331107	0.239431	0.156426
1.428155	0.454235	1.678942	0.749362	1.037226	2.769977	2.744425
1.969249	3.56071	1.982682	4.322052	5.523928	1.530836	2.092786
1.024378	2.561752	1.076142	3.333333	4.968035	0.468671	0.923361
0.83196	0.588235	0.951058	1.6875	2.081657	1.972308	2.012248
1.183644	0.127531	1.119147	1.297965	1.435156	2.229781	2.19798
0	1.541962	0.120214	2.315602	2.821869	0.906035	0.637418
0.194465	2.074978	0.115825	2.47619	4.158186	0.8102	0.488334
0.09057	1.424757	0	2.333864	2.93149	0.803299	0.611995

誤差は多いところで 5 % とかなりあった。簡略版といっても、さほど計算量が減るわけではないので、利用価値があるのか疑わしい。

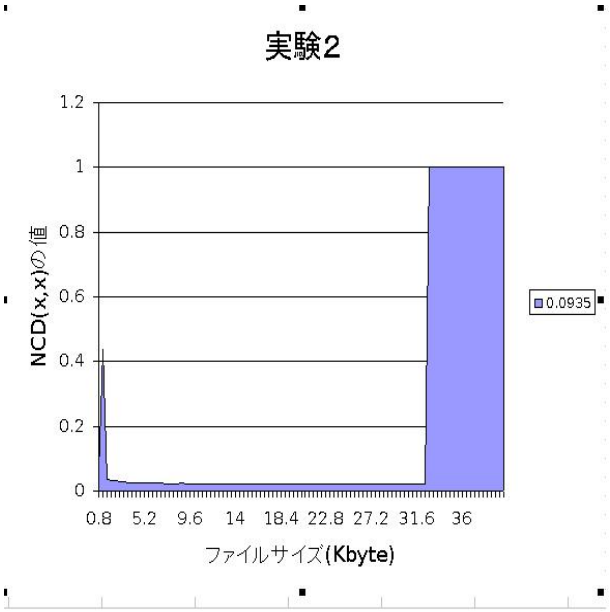
4.2 gzip における NCD(x,x) の限界を調べる

gzip は辞書式のアルゴリズムである。よって辞書サイズ内においては $C(xx) = C(x)$ をおおよそ満たすた。しかし、辞書のサイズには限界があり、ある一点を境に全く機能しなくなるのではないか。そこでファイルサイズを徐々に大きくしていき NCD(x,x) が 0 から離れるか。類似度判定における gzip の限界を探る。

実験

使用するファイル：ランダムの文字列からなるファイル (40Kbyte)

実験方法：ファイルの大きさを 400byte ずつ増やしていき、NCD を計算する。以下が結果のグラフである。



gzip の仕様書には 3 2 kbyte が辞書サイズと書いてあった。実験でも 32kbyte 付近を境に NCD の値が急激に増えている。

参考文献

[1] Ming Li, Member, IEEE, Xin Chen, Xin Li, Bin Ma, and Paul M. B. Vitanyi
The Similarity Metric (2004)

[2] 谷　聖一
谷聖一　データの複雑さ・データ表現の複雑さ

[3] Rudi Cilibrasi and Paul M.B. Vitanyi
Clustering by Compression (2005)