Uses of Information Theory in Medical Imaging

Norbert Schuff, Ph.D.
Center for Imaging of Neurodegenerative Diseases
Norbert.schuff@ucsf.edu
Measurement & Information

Objective

Learn how to quantify information
Topics

I. Measurement of information
II. Image registration using information theory
III. Imaging feature selection using information theory
IV. Image classification based on information theoretic measures
Information and Uncertainty

Random Generator

\[ \text{Which char comes next?} \]

\[ \Rightarrow \text{AAAAAAA} \quad \text{A} \quad \log(1) = 0 \]

\[ \Rightarrow \text{ABACCCACB} \quad \text{A} \quad \log(1/3) = -\log(3) \]

\[ \Rightarrow 011000101 \quad 1 \quad \log(1/2) = -\log(2) \]

Both combined:

\[ -\log(3) - \log(2) = -\log(6) \]

Note, we assume each symbol is likely to appear at equal chance.
More On Uncertainty

Random Generator Of M Symbols

Which char comes next?

→ AB@CBCA  @

Some symbols may appear more frequent than others

Entropy: H

\[-p_i \log_2 (p_i)\]

Many: \( H_{Shannon} = - \sum_{i=1}^{M} p_i \log_2 (p_i) \)

\[\sum_{i=1}^{M} p_i = 1\]
Example: Entropy Of mRNA

...ACGTAACACCACCTG

Assume “A” occurs 50% of all instances, “C” = 25%, “G” and “T”, each = 12.5%

\[ p_A = \frac{1}{2}; \ p_C = \frac{1}{4}; \ p_G = \frac{1}{8}; \ p_T = \frac{1}{8} \]

Logodds.. \(- \log_2 (p_i)\): \( p_A = 1; \ p_C = 2; \ p_G = 3; \ p_T = 3 \)

Entropy(bits) \( H = \frac{1}{2} \cdot 1 + \frac{1}{4} \cdot 2 + \frac{1}{8} \cdot 3 + \frac{1}{8} \cdot 3 = 1.75 \)
Concept Of Entropy

• Shannon Entropy formulated by Claude Shannon

• American mathematician
• Founded information theory with one landmark paper in 1948
• Worked with Alan Turin on cryptography during WWII

• History of Entropy
• 1854 – Rudolf Clausius, German physicist, developed theory of heat
• 1876 – Williard Gibbs, American physicist, used it for theory of energy
• 1877 – Ludwig Boltzmann, Austrian physicist, formulated theory of thermodynamics
• 1879 – Gibbs re-formulated entropy in terms of statistical mechanics
• 1948 - Shannon
Three Interpretations of Entropy

• The uncertainty in the outcome of an event
  – Systems with frequent event have less entropy than systems with infrequent events.

• The amount of surprise (information)
  – An infrequently occurring event is a greater surprise, i.e. carries more information and thus has higher entropy, than a frequently occurring event.

• The dispersion in the probability distribution
  – An uniform image has a less disperse histogram and thus lower entropy than a heterogeneous image.
Generalized Entropy

• The following generating function can be used as an abstract definition of entropy:

\[
H(P) = h \left( \frac{\sum_{i=1}^{M} v_i \cdot \varphi_1(p_i)}{\sum_{i=1}^{M} v_i \cdot \varphi_2(p_i)} \right)
\]

• Various definitions of these parameters provide different definitions of entropy.
  – Found over 20 definitions of entropy
Various Formulations Of Entropy

<table>
<thead>
<tr>
<th>Names</th>
<th>$h(x)$</th>
<th>$\varphi_1(x)$</th>
<th>$\varphi_2(x)$</th>
<th>$v_i$</th>
</tr>
</thead>
<tbody>
<tr>
<td>Shannon</td>
<td>$x$</td>
<td>$-x \log x$</td>
<td>$x$</td>
<td>$v$</td>
</tr>
<tr>
<td>Renyi</td>
<td>$(1 - r)^{-1} \log x$</td>
<td>$x^r$</td>
<td>$x^r$</td>
<td>$v$</td>
</tr>
<tr>
<td>Aczel</td>
<td>$x$</td>
<td>$-x^r \log x$</td>
<td>$x^r$</td>
<td>$v$</td>
</tr>
<tr>
<td>Aczel</td>
<td>$(s - r)^{-1} \log x$</td>
<td>$x^r$</td>
<td>$x^s$</td>
<td>$v$</td>
</tr>
<tr>
<td>Aczel</td>
<td>$(1/s) \arctan x$</td>
<td>$x^r \sin(s \log x)$</td>
<td>$x^r \cos(s \log x)$</td>
<td>$v$</td>
</tr>
<tr>
<td>Varma</td>
<td>$(m - r)^{-1} \log x$</td>
<td>$x^r - m + 1$</td>
<td>$x^r$</td>
<td>$v$</td>
</tr>
<tr>
<td>Varma</td>
<td>$(m(m - r))^{-1} \log x$</td>
<td>$x^r/m$</td>
<td>$x^r$</td>
<td>$v$</td>
</tr>
<tr>
<td>Kapur</td>
<td>$(1 - t)^{-1} \log x$</td>
<td>$x^{t+s-1}$</td>
<td>$x^s$</td>
<td>$v$</td>
</tr>
<tr>
<td>Hadra</td>
<td>$(1 - s)^{-1}(x - 1)$</td>
<td>$x^s$</td>
<td>$x^s$</td>
<td>$v$</td>
</tr>
<tr>
<td>Arimoto</td>
<td>$(t - 1)^{-1}(x^t - 1)$</td>
<td>$x^{1/t}$</td>
<td>$x^{1/t}$</td>
<td>$v$</td>
</tr>
<tr>
<td>Sharma</td>
<td>$(1 - s)^{-1}(e^x - 1)$</td>
<td>$(s - 1)x \log x$</td>
<td>$(s - 1)x \log x$</td>
<td>$v$</td>
</tr>
<tr>
<td>Sharma</td>
<td>$(1 - s)^{-1}(x^{s-1} - 1)$</td>
<td>$x^r$</td>
<td>$x^r$</td>
<td>$v$</td>
</tr>
</tbody>
</table>
## Various Formulations of Entropy II

<table>
<thead>
<tr>
<th>Names</th>
<th>( h(x) )</th>
<th>( \varphi_1(x) )</th>
<th>( \varphi_2(x) )</th>
<th>( v_i )</th>
</tr>
</thead>
<tbody>
<tr>
<td>Taneja</td>
<td>(x)</td>
<td>(-x^r \log x)</td>
<td>(x)</td>
<td>(v)</td>
</tr>
<tr>
<td>Sharma</td>
<td>((s-r)^{-1}x)</td>
<td>(x^r - x^s)</td>
<td>(x)</td>
<td>(v)</td>
</tr>
<tr>
<td>Sant'Anna</td>
<td>((\sin s)^{-1}x)</td>
<td>(-x^r \sin(s \log x))</td>
<td>(x)</td>
<td>(v)</td>
</tr>
<tr>
<td>Sant'Anna</td>
<td>(\left(1 + \frac{1}{\lambda}\right) \log(1 + \lambda) - \frac{x}{\lambda})</td>
<td>((1 + \lambda x) \log(1 + \lambda x))</td>
<td>(x)</td>
<td>(v)</td>
</tr>
<tr>
<td>Picard</td>
<td>(x)</td>
<td>(-x \log \left(\frac{\sin(sx)}{2 \sin(s/2)}\right))</td>
<td>(x)</td>
<td>(v)</td>
</tr>
<tr>
<td>Picard</td>
<td>(x)</td>
<td>(-\log x)</td>
<td>(1)</td>
<td>(v_i)</td>
</tr>
<tr>
<td>Picard</td>
<td>((1 - r)^{-1} \log x)</td>
<td>(-\log x)</td>
<td>(1)</td>
<td>(v_i)</td>
</tr>
<tr>
<td>Picard</td>
<td>((1 - s)^{-1}(e^x - 1))</td>
<td>((s - 1) \log x)</td>
<td>(1)</td>
<td>(v_i)</td>
</tr>
<tr>
<td>Picard</td>
<td>((1 - s)^{-1}(x^{s-1} - 1))</td>
<td>((s - 1) \log x)</td>
<td>(1)</td>
<td>(v_i)</td>
</tr>
</tbody>
</table>
Entropy In Medical Imaging

Averaging Map Histogram Entropy

N=1

H=6.0 bits

N=40

H= 4.0

Histograms image brain only Co-variance map

H= 3.9

H= 10.1
Entropy Of Noisy Images

Signal-to-Noise Level

SNR=20

SNR=2

SNR=0.2

Entropy Map

Decreasing SNR Levels

Entropy vs. SNR
Image Registration Using Information Theory

Objective

Learn how to use Information Theory for imaging registration and similar procedures
Image Registration

• Define a transform $T$ that maps one image onto another image such that some measure of overlap is maximized.
  
  – Discuss information theory as means for generating measures to be maximized over sets of transforms.
Entropy In Image Registration

• Define estimate of joint probability distribution of images:

  – 2-D histogram where each axis designates the number of possible intensity values in corresponding image

  – each histogram cell is incremented each time a pair \((I_1(x,y), I_2(x,y))\) occurs in the pair of images (“co-occurrence”)

  – if images are perfectly aligned then the histogram is highly focused; as the images become mis-aligned the dispersion grows

  – recall one interpretation of entropy is as a measure of histogram dispersion
Joint Entropy

• Joint entropy (entropy of 2-D histogram):

\[ H(X,Y) = - \sum_{x_i \in X, y_i \in Y} p(x_i, y_i) \cdot \log_2[p(x_i, y_i)] \]

• \( H(X,Y) = H(X) + H(Y) \) only if \( X \) and \( Y \) are completely independent.

• Image registration can be guided by minimizing joint entropy \( H(X,Y) \), i.e. dispersion in the joint histogram for images is minimized
Example

Joint Entropy of 2-D Histogram for rotation of image with respect to itself of 0, 2, 5, and 10 degrees
Entire Rotation Series

Rotation (degrees)

0  2  5  10

3.82  6.79  6.98  7.15
Mutual Information

\[ MI(X, Y) = \sum_{x_i \in X} \sum_{y_i \in Y} p(x_i, y_i) \cdot \log \left( \frac{p(x_i, y_i)}{p(x_i) \cdot p(y_i)} \right) \]

- Measures the dependence of the two distributions
- \( MI(X, Y) \) is maximized when the images are co-registered

- In feature selection choose the features that minimize \( MI(X, Y) \) to ensure they are not related
Use Of Mutual Information for Image Registration

- Alternative definition(s) of mutual information MI:
  - $\text{MI}(X,Y) = H(Y) - H(Y|X) = H(X) - H(X|Y)$
    - amount by which uncertainty of $X$ is reduced if $Y$ is known or vice versa.
  - $\text{MI}(X,Y) = H(X) + H(Y) - H(X,Y)$
    - maximizing $\text{MI}(X,Y)$ is equivalent to minimizing joint entropy $H(X,Y)$

- Advantage in using mutual information over joint entropy is that MI includes the entropy of each distribution separately.

- Mutual information works better than simply joint entropy in regions with low contrast where there will be high joint entropy but this is offset by high individual entropies as well - so the overall mutual information will be low

- Mutual information is maximized for registered images
Properties of Mutual Information

• MI is symmetric:
  \[ \text{MI} (X,Y) = \text{MI}(Y,X) \]

• \( \text{MI}(X,X) = H(X) \)

• \( \text{MI}(X,Y) \leq H(X) \)
  – Information each image contains about the other image cannot be greater than the total information in each image.

• \( \text{MI}(X,Y) \geq 0 \)
  – Cannot increase uncertainty in X by knowing Y

• \( \text{MI}(X,Y) = 0 \) only if X and Y are independent
Processing Flow for Image Registration Using Mutual Information

1. Input Images
2. Pre-processing
3. Probability Density Estimation
4. Mutual Info Estimation
5. Image Transformation
6. Optimization Scheme
7. Output Image
Relative Entropy

- Assume we know $X$ then we seek the remaining entropy of $Y$

$$H(Y \mid X)$$

$$= - \sum_{x_i \in X} p(x_i) \sum_{y_i \in Y} p(y_i \mid x_i) \cdot \log[p(y_i \mid x_i)]$$

$$= - \sum_{x_i \in X, y_i \in Y} p(x_i, y_i) \log \left[ \frac{p(y_i, x_i)}{p(x_i)} \right]$$

Also known as
Kullback-Leibler divergence;
An index of information gain
Measurement Of Similarity Between Distributions

- Question: How close (in bits) is a distribution $X$ to a model distribution $\Omega$?

$$D(X \parallel \Omega)_{KL} = \sum_{x_i \in \Omega} p(x_i) \cdot \log \left( \frac{p(x_i)}{q(x_i)} \right)$$

Relative Entropy

Kullback-Leibler Divergence

- $D_{KL} = 0$ only if $X$ and $\Omega$ are identical; otherwise $D_{KL} > 0$
- $D_{KL}(X \parallel \Omega)$ is NOT equal to $D_{KL}(\Omega \parallel X)$

- Symmetrical Form

$$sD(X \parallel \Omega)_{sKL} = \frac{1}{2} \left( D(X \parallel \Omega) + D(\Omega \parallel X) \right)$$
Uses Of The Symmetrical KL

How much more noisy is image A compared to image B?

Map

Histograms

sKL distance B from A

Increasing noise level
Uses Of The Symmetrical KL

Detect outlier in a image series

![Image showing series of images and a graph representing sKL vs Series Number.](image)
Image Feature Selection Using Information Theory

Objective

Learn how to use information theory for determining imaging features
Mutual Information based Feature Selection Method

- We test the ability to separate two classes (features).

- Let $X$ be the feature vector, e.g. co-occurrences of intensities.

- $Y$ is the classification, e.g. particular intensities.

- How to maximize the separability of the classification?
Mutual Information based Feature Selection Method

• MI tests a feature’s ability to separate two classes.
  – Based on definition 3) for mutual information

\[
MI(X,Y) = \sum_{x_i \in X} \sum_{y_i \in Y} p(x_i, y_i) \cdot \log \left( \frac{p(x_i, y_i)}{p(x_i)p(y_i)} \right)
\]

  – Here X is the feature vector and Y is the classification
    • Note that X is continuous while Y is discrete
  – By maximizing \(MI(X,Y)\), we maximize the separability of the feature
    • Note this method only tests each feature individually
Joint Mutual Information based Feature Selection Method

- Joint MI tests a feature’s independence from all other features:

- Two implementations:
  - 1) Compute all individual MI.s and sort from high to low
  - 2) Test the joint MI of current feature while keeping all others
    - Keep the features with the lowest JMI (implies independence)
    - Implement by selecting features that maximize:
Information Of Time Series

Objective

Learn how to use information theory for quantifying functional brain signals
How much information carry FMRI fluctuations?

Conventional Analysis – Correlations, e.g.:

\[
\text{corr}(A, B) = \frac{\sum_{i=1}^{M} (a_i - \bar{a}) \cdot (b_i - \bar{b})}{(M - 1) \cdot \text{std}(a) \cdot \text{std}(b)}
\]

\(<\text{Corr}\rangle\) does not measure information, aka uncertainty.
Quantification of Uncertainty in fMRI using Entropy

Block Entropy $H(L)$:
Uncertainty in identifying fluctuation patterns.
Measuring Difficulty in Overcoming Uncertainty

- **Block Entropy** $H(L)$: uncertainty in identifying fluctuation patterns (e.g., patterns of block size $L$).
- **Excess Entropy** $EE$: intrinsic pattern memory
- **Transient Information** $TI$: $TI = \sum_L EE - H(L)$
Transient Information in fMRI

Figure 2. Transient information maps of three different subjects, with different number of measurements (L=1,3,10)
Image Classification Based On Information Theoretic Measures

Objective

Learn how to use information theory for classification of complex image patterns
Complexity In Medical Imaging

Complexity

• In imaging, local correlations may introduce uncertainty in predicting intensity patterns.

• Complexity describes the inherent difficulty in reducing uncertainty in predicting patterns.

• Need a metric to quantify complexity, i.e. reduce uncertainty in the distribution of observed patterns.
Example: Patterns Of Cortical Thinning

Brain MRI

Histological Cut

Cortical ribbon
Patterns Of Cortical Thinning In Dementia

Alzheimer’s Disease

Frontotemporal Dementia

Warmer colors represent more thinning
Information Theoretic Measures of Image Complexity

Image Segment

Parse window size L over image

Information theoretic quantification

Three states: \( Z_i = \square \quad \square \quad \blacksquare \)

Entropy:
\[
H = - \sum_{Z, Z'} P(Z, Z') \cdot \log(P(Z, Z'))
\]

Statistical Complexity:
\[
SC = - \sum_{Z} \left( \sum_{Z'} P(Z | Z') \cdot \log(P(Z | Z')) \right)
\]

Excess Entropy:
\[
EE = - \sum_{L} H(L)
\]
Summary Complexity Measures

- **Entropy (H)** – measures number and uniformity of distributions over observed patterns (joint entropy). For example, a higher H represents increasing spatial correlations in image regions.

- **Statistical Complexity (SC)** – quantifies the information, i.e. uncertainty, contained in the distribution of observed patterns. For example, a higher SC represents an increase in locally correlated patterns.

- **Excess Entropy (EE)** – measures convergence rate of entropy. A higher EE represents increase long range correlations across regions.
Complexity Based Detection Of Cortical Thinning (Simulations)

H, SC, and EE automatically identify cortical thinning and do a good job at separating the groups for which the cortex was thinned from the group for which there was no thinning.
Detection Of Cortical Thinning Pattern In Dementias Using Complexity Measures

RGB representation: H; SC; EE
More saturated red means more spatial correlations;
More saturation of blue means more locally correlated patterns;
More saturated green means more long range spatial correlations;
Simultaneous increase/decrease of H, EE, SC results in brighter/darker levels of gray
<table>
<thead>
<tr>
<th>Metric</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>Entropy</td>
<td>Amount of uncertainty in observations</td>
</tr>
<tr>
<td>Conditional entropy</td>
<td>Uncertainty in observations given the distribution of other observations`</td>
</tr>
<tr>
<td>Joint entropy</td>
<td>Uncertainty in observations across multiple distributions</td>
</tr>
<tr>
<td>Relative entropy, aka Kullback-Leibler divergence</td>
<td>Similarity between distributions</td>
</tr>
<tr>
<td>Mutual Information</td>
<td>The amount of overlap between distributions</td>
</tr>
<tr>
<td>Statistical complexity</td>
<td>Uncertainty in linked conditional distributions</td>
</tr>
<tr>
<td>Excess entropy</td>
<td>Convergence of entropy</td>
</tr>
<tr>
<td>Transient information</td>
<td>Difficulty in reducing uncertainty</td>
</tr>
</tbody>
</table>
Literature

Author: James P Sethna

Author: H Haken