Summarizing Complexity in High Dimensional Spaces

Karl Young (karl.young@ucsf.edu) – *University of California, San Francisco,* USA

As the need to analyze high dimensional, multi-spectral data on complex physical systems becomes more common, the value of methods that glean useful summary information from the data increases. This paper describes a method that uses information theoretic based complexity estimation measures to provide diagnostic summary information from medical images. Implementation of the method would have been difficult if not impossible for a non expert programmer without access to the powerful array processing capabilities provided by SciPy.

Introduction

There is currently an explosion of data provided by high precision measurements in areas such as cosmology, astrophysics, high energy physics and medical imaging. When faced with analyzing such large amounts of high dimensional, multi-spectral data the challenge is to deduce summary information that provides physical insight into the behavior of the underlying system in a way that allows for generation and/or refinement of dynamical models.

A major issue facing those trying to analyze this type of data is the problem of dealing with a "large" number of dimensions both in the underlying index space (i.e. space or space-time) as well as the feature or spectral space of the data. Versions of the curse of dimensionality arise both from trying to generalize the methods of time series analysis to analysis in space and space-time as well as for data having a large number of attributes or features per observation.

It is here argued that information theoretic complexity measures such as those described in [Young1] can be used to generate summary information that characterizes fundamental properties of the dynamics of complex physical and biological systems. The interpretability and utility of this approach is demonstrated by the analysis of imaging studies of neurodegenerative disease in human brain. One important reason for considering such an approach is that data is often generated by a system that is non-stationary in space and/or time. This may be why statistical techniques of spatial image or pattern classification, that rely on assumptions of stationarity, have given inconsistent results when applied to magnetic resonance imaging (MRI) data. While various heuristic methods used for texture analysis have proven fruitful in particular cases of - for example - image classification, they typically do not generalize well or provide much physical insight into the dynamics of the system being analyzed. The methods described in this paper should be particularly effective in cases like classification of multi-spectral data from a particular class of physical object, i.e. for which the data to be analyzed and compared comes from a restricted class such as brain images from a set of subjects exhibiting the symptoms of one of a small class of neurodegenerative disease. The methods described allow for direct estimation of summary variables for use in classification of the behavior of physical systems, without requiring the explicit constructions described in [Crutch].

Methods

The complexity estimation methods used in this study were introduced in [Crutch] for time series analysis. The fundamental question addressed there was how much model complexity is required to optimally predict future values of a time series from past values. In addition a framework was provided for building the optimal model in the sense of being the minimally complex model required for reliable predictions.

Heuristic arguments and examples provided in [Young1] showed that only slight modifications were required to generalize the formalism for analysis of spatial data and in particular medical image data. Critical to the definition of complexity is the notion of "state" which provides the ability to predict observed values in a time series or image. Simply put, the complexity of the set of states required to describe a particular time series or image, is an indication of the complexity of the system that generated the time series or image. This in effect provides a linguistic description of the system dynamics by directly describing the structure required to infer which measurement sequences can be observed and which cannot. As described in [Crutch] to accurately and rigorously characterize the underlying complexity of a system, the set of states must in fact constitute a minimal set of optimally predictive states. How those criteria are defined and satisfied by the constructions outlined in this paper is described in [Young1], [Young2]. The simplest notion of complexity that arises from the above considerations involves a count of the number of states required for making optimal predictions. Since enumerated states can occur with varying frequencies during the course of observations, introducing the notion of state probability is natural. Shannon's original criteria for information [Shan], provides the simplest definition of an additive quantity associated with the probability distribution defined over a set of states. Complexity can then be described as an extensive quantity (i.e. a quantity that scales with measurement size) defined as the Shannon information of the probability distribution of the set of states describing the underlying system. For equally probable states this definition simply yields the log of the number of states as a measure of complexity. This notion of complexity, based on considerations of optimal prediction, is very different from the traditional

notion of Kolmogorov complexity [Cov], which quantifies series of random values as the most complex, based on considerations of incompressibility of a sequence. Here sequence is interpreted as a "program" in the context of computation, and data in the context of data analysis. Both notions of complexity provide important and complementary measures for characterizing structure in images. In the following, the optimal prediction based definition of complexity is the statistical complexity (SC) and the incompressibility based definition of complexity is entropy (H), since the Kolmogorov complexity corresponds, in the case of data analysis, to what physicists typically refer to as entropy [Cov]. A third quantity is excess entropy (EE), defined in [Feld]. EE is complementary to SC and H, and can be shown to provide important additional information. EE essentially describes the convergence rate of the entropy H to its asymptotic value as it is estimated over larger and larger volumes of the index space. The combination of EE, H and SC gives a robust characterization of the dynamics of a system.

The estimation and use of H, SC, and EE for classification of images, proceeds in 4 stages:

1. choice of an appropriate feature space (e.g. in a medical image analysis some combination of co-registered structural MRI, diffusion images,spectroscopic images, PET images, or other modalities).
2. segmentation (clustering) of feature space, i.e. clustering in the space of features without regard to coordinates (analogous to standard image segmentation for a single feature).
3. mapping of the clustered values back to the original coordinate grid and generation of estimates of H, SC, and EE from the image of clustered values.
4. classification of the data sets (e.g. images) based on the complexity estimates (e.g. via supervised or unsupervised learning algorithms)

The software implementation of the above methods is an open source package written in Python using SciPy and the Rpy [More] package to provide access to the statistical and graphical capabilities of the R statistical language [RDev] and supplemental libraries. The cluster and e1071 [Dimi] R packages were used for clustering and the AnalyzeFMRI [March] package for MR image processing. Image analysis was performed using this package on a 46 processor Beowulf cluster using the PyPAR [Niel] Python wrapper for the message passing interface (MPI). Complete (fully automated) processing of a single subject takes on the order of 40 minutes on a single 3 GHz processor.

Some important questions for future developments of the package include whether enough statistical capability will or should be provided directly in SciPy to obviate the need for inclusion of Rpy and R and how easy it will be to incorporate Ipython as a base platform for distributed processing.

In the next section I describe an illustrative analysis of structural MRI images from 23 cognitively normal (CN) subjects, 24 patients diagnosed with Alzheimer's disease (AD) and 19 patients diagnosed with frontal temporal dementia (FTD). The analysis and data are described in [Young3]. In brief: our feature space was the segmentation of each MRI image into gray matter, white matter and cerebrospinal fluid; we then applied a template of neighbouring voxels (2 neighboring voxels, compared to the next two voxels in the same line) to generate a local co-occurrence matrix of the three tissue classes, centered at each voxel; we applied the complexity metrics to this matrix, giving us an H, EE and SC measure at each voxel of each scan. We can then use regional or global summary statistics from these voxel-wise measures to classify scans according to diagnostic group.

Results

The variability of the three complexity measures in different brain regions is illustrated in Figure (1), separately for single representative CN, AD, and FTD subjects.

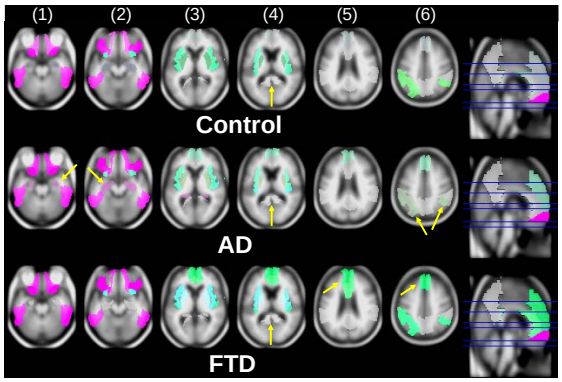

Figure 1

Simultaneous variability of entropy (H), excess entropy (EE) and statistical complexity (SC) of different brain regions in a single control subject, a single subject diagnosed with AD, and a single subject diagnosed with FTD, represented in an additive red-green-blue (RGB) color space.

An additive red-green-blue (RGB) color space is used to represent simultaneous values of H, EE, and SC. In this color space the value of H is represented on the red axis, EE on the green axis and SC on the blue axis. In this representation, a higher saturation of red represents a higher value of H, implying lack of correlation of structural patterns in an image region. Similarly, a higher saturation of green represents a higher value of EE, implying increased long range correlations of structural patterns and a higher saturation of blue represents a higher value of SC, implying an increase of locally correlated patterns. Accordingly, a simultaneous increase/decrease of all three complexity measures results in brighter/darker levels of gray. The most prominent effects in the AD subject compared to the CN and FTD subjects as seen in this representation are decreased correlation in the hippocampus (faint red regions, yellow arrows in columns 1 and 2) and diminished long range correlations of structural patterns in superior parietal lobe regions (faint green regions, arrows in column 6). In contrast, the most

prominent effect in the FTD subject compared with the CN and AD subjects is greater long range correlation in medial frontal lobe and anterior cingulum (intense green regions, arrows in columns 5 and 6).

An important practical question is whether the H, EE, and SC measures are able to distinguish between diagnostic groups as well as the current standard, which is to use local measures of cortical gray matter volume and thickness. In the following, we use logistical regression to classify scans, comparing performance using different measures.

Table (1) compares results using the structural complexity estimation against results on use of gray matter (GM) cortical thickness estimation using the FreeSurfer software on the same set of subjects.

Metric / groups	AD/CN (%)	FTD/CN (%)	AD/FTD (%)
Parietal GM volume	95 ± 4	81 ± 7	85 ± 6
Parietal GM thickness	96 ± 3	82 ± 6	86 ± 6
3 region complexity	92 ± 1	87 ± 1	91 ± 1

Table 1: logistical regression classification using Freesurfer and complexity metrics

In the table, comparisons are between classification accuracy based on structural complexity estimation and classification accuracy based on tissue volume and cortical thickness estimation (the parietal lobes provided the best separation between AD and CN subjects and the only significant separation between AD and FTD subjects for the volume and thickness estimates). For each, complexity or FreeSurfer, the regions providing the best separation between the groups are listed: for complexity the hippocampus, parietal lobe, precuneus, and Heschl's gyrus are taken together; for FreeSurfer we took measures of the thickness of parietal lobe gray matter (GM). This shows that structural complexity measures slightly outperformed volume and cortical thickness measures for the differential classification between AD and FTD as well as between FTD and CN. For the classification between AD and CN, volume and cortical thickness estimation achieved slightly higher classifications than structural complexity estimation.

Note that the classification results above may be close to the practical limit; clinician diagnosis of both AD and FTD does not have perfect agreement with post-mortem diagnosis from pathology, with errors in the same order as those reported here.

The results above compared pairwise between groups (CN vs AD, CN vs FTD, AD vs FTD). We can also assess prediction accuracy when trying to separate all three 3 groups at once, using linear discriminant analysis (LDA). This is illustrated graphically in Figures (2a), (2b), and (2c) which depict the projections onto the first two linear discriminants (labeled LD1 and LD2 in the Figures) from the LDA corresponding to the region selections for complexity estimation. This shows the expected result that group separation prominently increased with the use of focal measures, such as each of the 13 regions, as compared to global measures, such as whole brain.

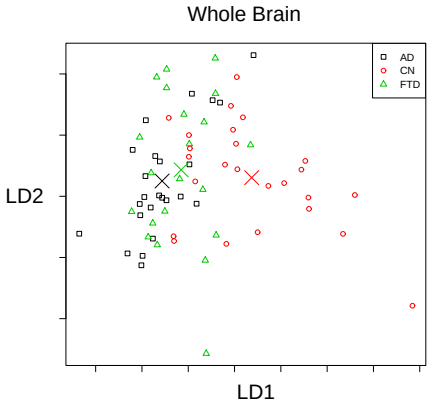

Figure 2 (a)

Results of linear discriminant analysis (LDA) using structural complexity estimates with x and y axes representing projections of complexity estimates onto the 1st and 2nd linear discriminants for the whole brain

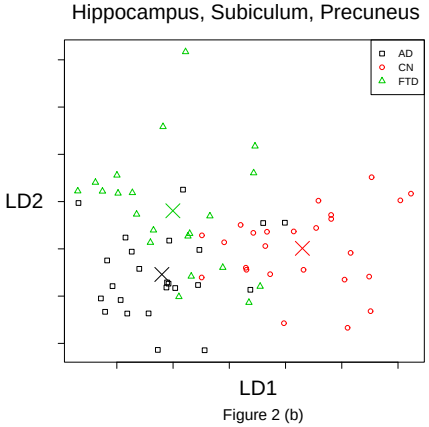

Figure 2 (b)

Results of linear discriminant analysis (LDA) using structural complexity estimates with x and y axes representing projections of complexity estimates onto the 1st and 2nd linear discriminants for the hippocampus, subiculum, and precuneus.

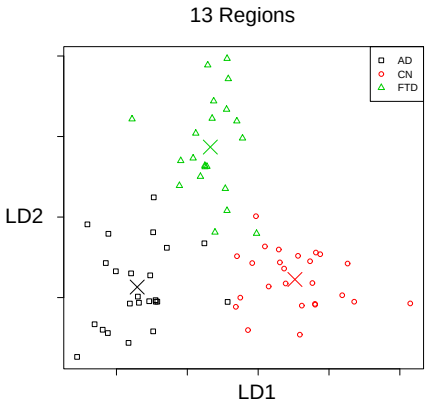

Figure 2 (c)

Results of linear discriminant analysis (LDA) using structural complexity estimates with x and y axes representing projections of complexity estimates onto the 1st and 2nd linear discriminants for all 13 regions.

Conclusion

This paper provides two main results. First, despite their simplicity and automated nature, use of structural complexity estimates is effective at capturing systematic differences on brain MRIs. They appear to be able to capture a variety of effects such as cortical volume loss and thinning. A second result is that complexity estimates can achieve similar classification separation between controls, AD and FTD patients, to that obtainable by highly specialized measures of cortical thinning. The classification accuracy provided by all of these methods is at or near the limit of the ability to reliably diagnose subjects during life, so further comparisons between methods will require improved clinical diagnosis, post-mortem diagnosis, or larger samples.

Though the complexity estimation results were promising, a number of issues remain before the methods can provide a concrete, interpretable tool suitable for clinical use. Future work will extend structural complexity estimation to multimodal imaging ([Young1]) in order to study neurodegenerative disease. This approach may be particularly effective as it does not depend on spatially confined effects in the different modalities for its classification power, as is the case for standard multivariate linear model image analysis. It also provides a more general and interpretable approach to understanding structural image properties than methods such as fractal and texture analysis.

Information theory based structural complexity estimation shows promise for use in the study and classification of large multivariate, multidimensional data sets including those encountered in imaging studies of neurodegenerative disease.

References

[Young1] Young K, Chen Y, Kornak J, Matson GB, Schuff N. Summarizing Complexity In High Dimensions. Physical Review Letters 2005; 94, 098701.

[Crutch] Crutchfield, JP, Young K. Inferring Statistical Complexity. Physical Review Letters 1989; 63, 105-107.

[Young2] Young K, Schuff N. Measuring Structural Complexity in Brain Images. Neuroimage 2008; 39(4):1721-30

[Shan] Shannon CE. A mathematical theory of communication. Bell Sys Tech Journal 1948; 27:379-423.

[Cov] Cover, T., Thomas J., 2006. Elements of Information Theory. Wiley-Interscience.

[Feld] Feldman DP, Crutchfield, JP. Structural Information in Two-Dimensional Patterns: Entropy Convergence and Excess Entropy. Physical Review E 2003; 67, 051104.

[More] Moreira, W., 2004. RPy Package. Available at http://rpy.sourceforge.net/

[RDev] R Development Core Team, 2004. R: A Language and Environment for Statistical Computing, R Foundation for Statistical Computing ISBN 3-900051-00-3. Available at http://www.r-project.org/

[Dimi] Dimitriadou, E., Hornik, K., Leisch, F., Meyer, D., Weingessel, A., 2004. R Package: e1071: Misc Functions of the Department of Statistics. Available at http://cran.r-project.org/

[March] Marchini, J.L., 2004.R Package: AnalyzeFMRI: Functions for Analysis of fMRI Datasets Stored in the ANALYZE Format. Availablae at http://cran.r-project.org/

[Niel] Nielsen O., Ciceri, G.P., Ramachandran, P., Orr, D., Kaukic, M., 2003. PyPAR - Parallel Python, Efficient and Scalable Parallelism Using the Message Passing Interface (MPI). Available at http://datamining.anu.edu.au/~ole/pypar/

[Young3] Young K, Du A, Kramer J, Rosen H, Miller B, Weiner M, Schuff N. Patterns of Structural Complexity in Alzheimer's Disease and Frontotemporal Dementia. Human Brain Mapping. In Press