Lecture 9 Structural Information Theory, III:

Structural Information Learning Model (SiL)

Angsheng Li

BeiHang University

19, Nov., 2019

Outline

1. The changing CS
2. The challenges
3. Structural information theory
4. Information processing (IP)
5. Learning model

The Changing Computing

What does a computer do?

1. The main stream of CS in the 20th century is to investigate:
   - 1.1 **Computable functions and computing devices**
   - 1.2 **Efficient algorithms**
2. The new mission of CS in the 21st century is to study:
   - 2.1 **Computing the real world**
   - 2.2 **Information processing** - to distinguish the laws from noises
   - 2.3 **Artificial intelligence?**

What Is Information Processing (IP)?

The challenges are:

1. What is the mathematical definition of **information processing**?

Definition

(Information processing)**Given a system S, the information processing for S is to distinguish the laws from the noises in S.**

2. What is the **mathematical theory** that supports the current information processing?

The mathematical theory about the limit of distinguishing the laws from noises in a system.

Big Data Phenomenon

1. There are relationships among individuals of data
   - How to build the system of big data (unstructured data)?
2. (Assumption) The laws of big data exist in the relationships of data
3. Big data is an observed system in which laws are embedded in a structure of noises
4. The mission of data analysis is hence:
   To distinguish the laws from noises
   This is
   To decode the laws from the observed system of data
   Grand challenge:
   What is the relationship between information and computation?

Shannon's Information

Shannon, 1948:
Given a distribution $p = (p_1, p_2, \cdots, p_n)$, the Shannon's entropy is

$$H(p) = -\sum_{i=1}^{n} p_i \cdot \log_2 p_i. \tag{1}$$

p_i is the probability that item i is chosen, $-\log_2 p_i$ is the "self-information" of item i.

- Shannon's information measure the uncertainty of a probabilistic distribution.
- This metric and the associated notions of noises form the foundation of information theory and the information theoretic study in all areas of the current science.
- Shannon's metric provides the foundation for the current generation information technology.

Principles of Communications

Shannon's theory perfectly solves the **fundamental questions of communications**.

- The lower bound of data compression - entropy
- The transmission rate of point to point communication: channel capacity
- Shannon's theory guarantees that lossless information transmission is possible

However, Shannon's theory **fails to support the current information processing**, in which we are asked to distinguish the laws from the noises from dataspace.
Noting that, data are observed from the real world, real world data is a system evolved in nature. It includes laws and noises. Laws are embedded in a system of noises.
Urgently call for an information theory that supports the information processing, that is a mathematical theory that distinguishes the laws from noises of an observed data-space.

Understanding of Information

1. Shannon entropy: $H(p)$ is the quantification of uncertainty contained in a probability distribution or random variable (essentially, a function)
2. **Information** is defined as the amount of uncertainty that is eliminated.
   Choosing an item i according to probability distribution p, we obtained an information of amount exactly $H(p)$.
3. This suggests that
   - 3.1 Entropy is a **static metric** associated with an object (probability distribution above)
   - 3.2 Information is a **dynamic metric** determined by both an **object** and an **action**, random selection here

Definition of Information -I

1. We define **entropy** to be the amount of uncertainty
2. Define **information** to be the amount of uncertainty that has been eliminated.

Definition of Information-II

- **Structural information theory** shows that for graphs G,
  entropy $\neq$ information
- **Information exists where entropy does.**
- **Rich information exists in complex systems**
- **Information exists everywhere!**

The Challenges

1. What is the entropy that is contained in a complex system such as graphs?
2. What is the quantification of information obtained from a complex system such as graphs?
3. **How to generate the maximum amount of information?**

Information vs Intelligence

1. Is information useful in computation?
2. What is the role of computation in information? Computing is decoding information, that is, eliminating uncertainty.
3. **What is intelligence?**
   Information is the basis of intelligence! Because, no information, then no intelligence.

Understanding of Intelligence

1. **To acquire information is to eliminate uncertainty!**
2. **How to eliminate the uncertainty in nature?**
3. **How to maximumly eliminate uncertainty?**

Encoding Tree

Definition

(Encoding tree of graphs) Let $G = (V, E)$ be an undirected and connected network. We define the *encoding tree T of G* as a tree T with the following properties:

(1) For the root node denoted λ, we define the set $T_\lambda = V$.

(2) For every node $\alpha \in T$, the immediate successors of α are $\alpha\hat{}\langle j\rangle$ for j from 1 to a natural number N ordered from left to right as j increases, where every internal node has at least two immediate successors. Therefore, $\alpha\hat{}\langle i\rangle$ is to the left of $\alpha\hat{}\langle j\rangle$ written as $\alpha\hat{}\langle i\rangle <_{\mathrm{L}} \alpha\hat{}\langle j\rangle$, if and only if $i < j$.

(3) For every $\alpha \in T$, there is a subset $T_\alpha \subset V$ that is associated with α.
For α and β, we use $\alpha \subset \beta$ to denote that α is an initial segment of β. For every node $\alpha \neq \lambda$, we use α^- to denote the longest initial segment of α, or the longest β such that $\beta \subset \alpha$.

Encoding Tree - II

(4) For every i, $\{T_\alpha \mid h(\alpha) = i\}$ is a partition of V, where $h(\alpha)$ is the height of α (note that the height of the root node λ is 0, and for every node $\alpha \neq \lambda$, $h(\alpha) = h(\alpha^-) + 1$).

(5) For every α, T_α is the union of T_β for all β's such that $\beta^- = \alpha$; thus, $T_\alpha = \cup_{\beta^- = \alpha} T_\beta$.

(6) For every leaf node α of T, T_α is a singleton; thus, T_α contains a single node of V.

Encoding Tree - III

(7) For every node $\alpha \in T$, if $T_\alpha = X$ for a set of vertices X, then we say that α is the **codeword** of X, and that X is the **marker** of α.

(8) For every vertex $v \in V$, there is a leaf node $\alpha \in T$ such that $T_\alpha = \{v\}$, that is, there is a unique **codeword** of v in T.

(9) Every leaf node in T is a codeword of a unique vertex (**marker**) in V.

Therefore, the set of the leaf nodes in T is the set of codewords of all the vertices in G.

Structural Entropy by an Encoding Tree

Definition

(Structural entropy of a graph by an encoding tree) For an undirected and connected network $G = (V, E)$, suppose that T is an encoding tree of G.

We define the **structural entropy of G by the encoding tree T** as follows:

$$\mathcal{H}^{T}(G) = -\sum_{\alpha \in T, \alpha \neq \lambda} \frac{g_\alpha}{2m} \log_2 \frac{V_\alpha}{V_{\alpha^-}}, \tag{2}$$

where g_α is the number of edges from nodes in T_α to nodes outside T_α, V_β is the volume of set T_β, namely, the sum of the degrees of all the nodes in T_β.

Structural Entropy

Definition

(Structural entropy) Let $G = (V, E)$ be a connected network.

- We define the **structural entropy of** G as follows:

$$\mathcal{H}(G) = \min_{T}\{\mathcal{H}^{T}(G)\}, \tag{3}$$

where T ranges over all of the encoding trees of G.

Information Compression and Enrichment

Let $G = (V, E)$ be a connected graph, and $\mathcal{P} = \{X_1, X_2, \cdots X_N\}$ be a partition of the vertices V.

Define **compression information of G by $\mathcal{P}$** as

$$\mathcal{C}^{\mathcal{P}}(G) = -\sum_{j=1}^{N} \frac{V_j - g_j}{V_j} \frac{V_j}{2m} \log_2 \frac{V_j}{2m}, \tag{4}$$

in which

- $\frac{V_j - g_j}{V_j}$ is the fraction of information of G compressed in X_j,
- $-\frac{V_j}{2m} \log_2 \frac{V_j}{2m}$ is the information of G contained in X_j,
- we call $-\frac{V_j - g_j}{V_j} \frac{V_j}{2m} \log_2 \frac{V_j}{2m}$ the **information enrichment** of X_j in G, written $\gamma_G(X_j)$.

We define the **information enrichment of G** by

$$\gamma(G) = \max_{X \subset V, V_X \leq m} \{\gamma_G(X)\}. \tag{5}$$

Information Enrichment vs Intelligence?

Project:

- Information enrichment of graphs
- Information distribution of graphs

Compressing Information by Encoding Tree

Definition

(Compressing information of a graph by an encoding tree) For an undirected and connected network $G = (V, E)$, suppose that T is an encoding tree of G.

We define the **compressing information of G by the encoding tree T** as follows:

$$\mathcal{C}^T(G) = -\sum_{\alpha \in T, \alpha \neq \lambda} \frac{V_\alpha - g_\alpha}{2m} \log_2 \frac{V_\alpha}{V_{\alpha^-}}, \tag{6}$$

where g_α is the number of edges from nodes in T_α to nodes outside T_α, V_β is the volume of set T_β, namely, the sum of the degrees of all the nodes in T_β.

Decoding Information by Encoding Tree

Definition

(Decoding information of a graph by an encoding tree) For an undirected and connected network $G = (V, E)$, suppose that T is an encoding tree of G.

We define the **decoding information of G by the encoding tree T** as follows:

$$\mathcal{D}^T(G) = \mathcal{H}^1(G) - \mathcal{H}^T(G). \tag{7}$$

Compressing Information

Definition

(Structural information) Let $G = (V, E)$ be a connected network.

- We define the **compressing information of** G as follows:

$$\mathcal{C}(G) = \max_{T}\{\mathcal{C}^{T}(G)\}, \tag{8}$$

where T ranges over all of the encoding trees of G.

Compressible Graphs

Definition

Given G, k and ρ, we say that G is (n, k, ρ)-compressible, if:

$$\rho^k(G) \geq \rho,$$

where

$$\rho^k(G) = \frac{\mathcal{C}^k(G)}{\mathcal{H}^1(G)}$$

Compressing and Decoding Principle

Theorem

Let $G = (V, E)$ be a connected network. Then:

$$\mathcal{C}(G) = \mathcal{H}^1(G) - \mathcal{H}(G) = \mathcal{D}(G). \tag{9}$$

Therefore, **any information lost in the compression of data can be losslessly decoded by an encoding tree, the decoder**.

This means that

For either unstructured or structured data, data compression will never loss any information

Shannon Entropy and Structural Entropy Are Complement

The compressing and decoding theorem:

$$\mathcal{C}(G) = \mathcal{H}^1(G) - \mathcal{H}(G) = \mathcal{D}(G). \tag{10}$$

indicates that the Shannon entropy minus the structural entropy is exactly the compressing information and the decoding information, so that

Shannon entropy and the structural entropy combining together characterise the information lost in data compression and the information recovered from a structural decoder.

Shannon Entropy vs Structural Entropy

- Metric
  Shannon: measure as a number, Structural Entropy: Measure as a number with an accompanying encoding tree that determines an encoding minimising uncertainty
- Objects
  Shannon: unstructured objects, Structural entropy: both structured and unstructured objects
- Dimensionality
  Shannon: One- dimensional, Structural entropy: High dimensional
- Measuring methods
  Shannon: Global, SE: Both local and global
- Role
  S: measuring uncertainty, SE: Simultaneously measuring uncertainty and decoding the essential structure

Shannon Entropy vs Structural Entropy- Big Picture

- Shannon Information Theory
  A mathematical theory of communication - point-to-point transmission
- Structural Information Theory
  A mathematical theory of information processing (IP)

Theory

- Structural Information Theory
- Information Theoretical Approach to Graphs
- Information Optimization Theory
- Network Theory
- Coding Theory of Big Data

A New Theory for Information Processing

Provable Theory for Data Analysis

Principles of Structural Information Theory

- **Structural entropy minimisation** is the principle for data analysis
- **Encoding tree** optimizes data structure of massive data
- **Encoding** eliminates the uncertainty that is embedded in a complex system
- **Structural information decoding** is a general model for information processing
- **Information is generated by structures**, providing a new understanding of the notion of information

The Role of Structure in Information Theory

What we have learnt from the structural information theory are:

- **Structure** plays a key role in information theory
- **Structural information** finds the encoding tree that minimizes the uncertainty (or random variations) in a system
- However, **randomness or noises** play a key role in learning and game
  This is the difference between structural information and learning/game
- **A natural question: Does structure play a role in AI and game?**

Structural Information Theoretical Foundations of Artificial Intelligence

Observations:

1. Intelligence generates in the form of social groups
2. The structure of social groups plays a key role in the generation of intelligence
3. Some structures generate intelligence, but some others fail

Big challenge: **What is the structural generating theory of intelligence**?

Information vs Intelligence

Decoding information is the amount of uncertainty we can reduce, we are intelligent people.

Big challenge:

Is there an information theoretical definition of intelligence?

What is the relationship between information and intelligence?

Structural Information Theoretical Approach to Machine Learning

Statistical learning is perhaps the most successful part of machine learning. However, statistics works only on unstructured probability distributions.
If information is key to learning, then there should be an information theoretical learning theory.

Information must be a key to learning! Why?

Big challenge: **Is there a new structural information learning theory?**

Game vs Information

Information must be the basis of game. We have shown that structure and randomness are key to the generation of information. The fundamental questions are hence:

1. What are the roles of structure and randomness in game?
2. What is the role of information in game
3. Is there an information theoretical direction of game theory?
4. **What is the role of game in intelligence?**

Structural Information Learning Machine (SiLM)

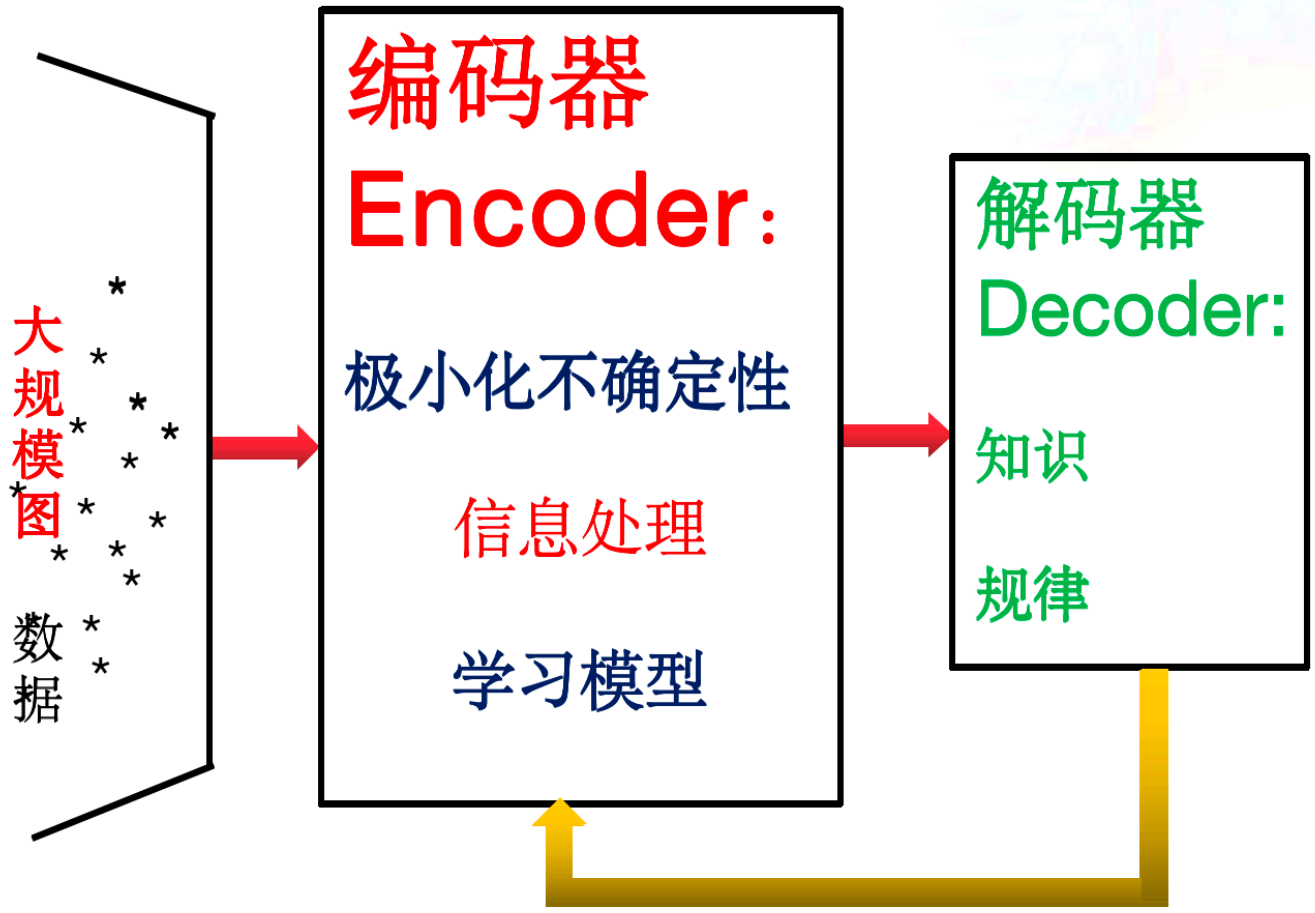

Figure: Learning Model.

Human-Like Learning?

1. Children learn - I want, to distinguish oneself from outside world
2. To understand the **certainty and uncertainty** of oneself
3. To understand the **certainty and uncertainty** of others
4. To understand the **certainty and uncertainty** of the environment
5. **To build knowledge, and to extract laws from knowledge**
6. **To use knowledge and laws to encode future knowledge**
7. **To create?**: Game could be the motivation of creation, providing a new mechanism of intelligence

An information theoretical understanding!

Encoding Tree of Graphs

A graph can be regarded as a structured dataset, in which there is an essential structure that supports the semantics of the dataset.
The encoding tree of the graph that **minimises the uncertainty** of the graph is defined as the essential structure of the graph. The essential structure represents the **knowledge structure** of the graph.

Encoding Tree=Knowledge

This suggests the **encoding tree representation of knowledge**.

The Laws

The **laws** of a graph is defined as the **rules** that generate the knowledge of the graph, i.e., the encoding tree of the graph. From the knowledge to laws, there is a procedure of **abstraction**, which extracts the laws from the encoding tree.

Decoding

The **Learning** proceeds as follows:

1. Structural entropy minimisation **determines and decodes** an encoding tree of the graph that minimises the uncertainty occurred in the graph.
2. **Extracts** the laws from the encoding tree or knowledge structure of the graph.

This procedure **acquires** the **knowledge and laws** of a structured dataset by an encoding that minimises the uncertainty.

Encoding with Knowledge and Laws

Suppose that we have an encoding tree T of an observed graph G that minimises the uncertainty occurred in G and that there are more new data D connects to the dataset in G. The structural entropy minimisation allows us to encode the new data D into T to form a new encoding tree T'. In the encoding of D into T, we are allowed to use even the laws L, written, of T. This means that the learning from D has already used the knowledge and laws of G.

This procedures provides a dynamical evolution of learning from the newly observed dataset by using the acquired knowledge and laws.

Clearly, the **learning** process is a procedure of:

Encoding + Decoding

Learning the Structure from Unstructured Dataset

Suppose that we are given observed **unstructured dataset** D.
We know that the knowledge and laws of D exist in the relationships among the data in the dataset.
To extract the knowledge and laws of D, we first define the measure of relationship between any two data points x and y in D.
Usually, there are several ways to define the measure of relationship between x and y. For each such a way, we form a structured dataset G_D. This provides several graphs $G_1, G_2, \cdots, G_N$ say, each corresponds to a measurement of the relationships between the data points in D.

Measurements of Relationship of Data Points

The structural entropy minimisation allows us to choose the i that minimises the one-dimensional structural entropy, that is,

$$\mathcal{H}^1(\mathrm{G_i}) = \min_{\mathrm{j}}\{\mathcal{H}^1(\mathrm{G_j})\} \tag{11}$$

Then, G_i is the preferred structured data set of D.
Let $G = G_i$.

Noise Amplifier

In G, the weight $w(e)$ for each e is usually noisy. We introduce a **noise amplifier** to remove the noises. It proceeds as follows:

1. Introduce a **noise amplifier** σ
2. For every edge e in G, set
   – $w(e) \leftarrow w(e) + \sigma$
   – We use G^σ to denote the graph obtained
3. Let σ be

$$\min \mathcal{H}^1(G^\sigma)$$

4. Set
   – $G' \leftarrow G^\sigma$

Sparsification

G' is usually dense. We will construct a graph $G^\star$ from G' such that the important edges are kept, and the trivial or noisy edges are removed. For this, we introduce a **sparsification operator**.
As usual, there are many ways of the sparsification. Structural entropy minimisation allows us to choose the sparsification that minises the uncertainty of the graph.
This constructs a graph $G^\star$ from G by using the sparsification. Let $G = G^\star$. Then the encoding tree of G is the encoding of the unstructured dataset D, providing the knowledge of D.

Principles of Learning

The learning machine satisfies the following properties:

1. The learning is the **encoding** that minimises the uncertainty occurred in the dataset
2. The principle of the learning is **to minimise** the uncertainty occurred in the dataset
3. The **learning machine** is **to minimise the uncertainty by encoding** and simultaneously **by using the knowledge** and even **laws** learnt previously. This property ensures that our learning machine can actually realise **abstracting**.
4. This learning process is highly similar to the **human-learning**

The High Level Overview of the Learning Machine

The high level overview:

1. **To learn is to minimise the uncertainty**
2. **Knowledge and laws exist in the relationship among the data points**
3. **Knowledge and laws minimise the uncertainty occurred in the dataset**
4. **Knowledge and laws are achieved by encoding and decoding**
5. **Knowledge and laws can be used in encoding and decoding**

Data Collection

The input of our learning machine is the dataset observed from the real world.
The dataset contains all the observations of the real world.

The Hypothesis Space

Our learning machine assumes that:

1. 事物由必然性和偶然性构成
2. 规律就是必然性
3. 知识就是不确定性最小化的数据集的编码树
4. The **hypothesis space** is the set of encoding trees. Therefore, the goal of learning is to find the **encoding tree** that minimises the uncertainty
   This is different from the existing models with hypothesis space to be a set of functions or probability distributions.

Structural Information Learning Theory

- The mathematical limits
- Characterisation of the learnable and the unlearnable

Realising Inductive Learning

Structural Information Learning vs **Learning from examples**

Structural Information Learning vs Deep Learning

Structural information principle for deep learning?

Structural Information Learning vs Statistical Learning

The relationship between

Minimizing uncertainty vs

Sampling

Structural Information Learning of Bayesian and Neural Networks

Structural information theoretical understanding of Bayesian networks and neural networks

Modeling the Environment

Environment consists of:

1. Structure
2. Random variations and/or noises

This understanding is different from that of being a probability distribution.

Definition of Intelligence

Definition

Intelligence is the **information** acquired from the dataset observed from the real world.

Grand Challenges of Structural Information Learning (SiL)

- Grand applications such as in biological data analysis, financial data analysis etc
- Mathematical definition of intelligence
- Generating models of intelligence

References

1. A. Li, Y. Pan, Structural Information and Dynamical Complexity of Networks, IEEE Transactions on Information Theory, **62**, No. 6, pp. 3290 - 3339, 2016.
2. Brooks, F. P., Jr. Three great challenges for half-century-old computer science. Journal of the ACM, **50** (1), pp 25 - 26 (2003).
3. C. Shannon, The lattice of information, IEEE Transactions on Information Theory, **1**, No. 1, pp. 105 - 107, 1953.
4. Angsheng Li, Xianchen Yin, Bingxiang Xu, Danyang Wang, Jimin Han, Yi Wei, Yun Deng, Ying Xiong, Zhihua Zhang, Decoding topologically associating domains with ultra-low resolution Hi-C data by graph structural entropy, NATURE COMMUNICATIONS,(2018) 9:3265.

Thank You!