Individual and Collective Information for Generating Interpretable Models of Multi-Layered Neural Networks

Ryotaro Kamimura

Kumamoto Drone Technology and Development Foundation

Techno Research Park, Techno Lab 203

1155-12 Tabaru Shimomashiki-Gun Kumamoto 861-2202

IT Education Center

Tokai University

Japan

ryotarokami@gmail.com

Outline

- **Problems of interpretation**
  - Understanding cognitive functions, Critical decision making, improving generalization
  - Difficulty to interpret the main mechanism of neural networks
- **Collective interpretation**
  - Interpreting neural networks, considering all possible representations generated by learning against conditional, individual and intuitive interpretation
- **Network compression**
  - For easy interpretation, we compress multi-layered numeral networks into the simplest ones without hidden layers
  - In addition, partial compression is used to see the state of intermediate layers
- **Combination of Individual and Collective Information**
  - To control information more flexibly for better interpretation, we introduce two types of information, namely, individual and collective information, and control them flexiblty
- **Application to bankruptcy data set**
  - We could detect **linear and independent relations** between inputs and outputs, hidden in complicated non-linear relations

Problems of Interpretation

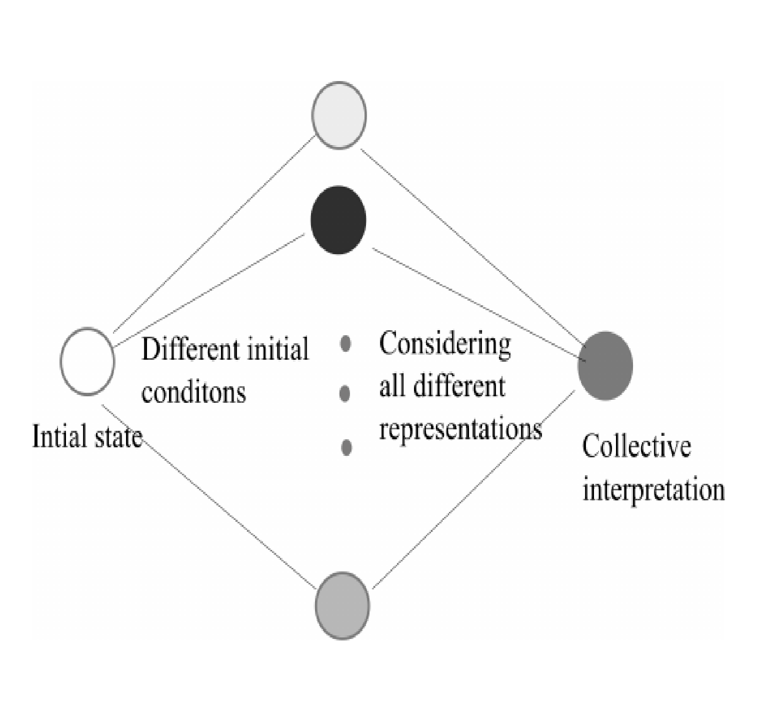
- **Multi-layered neural networks**
  - Applied to many application areas with good performance in improving generalization
- **Problem of interpretation**
  - **Complexity**
    - As the complexity becomes larger, the interpretation of neural networks has become a very serious problem
-
  Necessity of Interpretation
  - **Research Objective**
    - The objective of neural networks is to understand how human cognitive functions work by simulating them. Thus, it is necessary understand the inference mechanism of neural networks
  - **Critical decision making**
    - For application areas with critical decision making such as medical and business applications, it is absolutely necessary to interpret and explain the inference mechanism of neural networks
  - **Improving generalization**
    - In addition, to improve the general performance of neural networks, we need to understand how neural networks respond to inputs to produce outputs

Many Problems of Conventional Methods for Interpretation

- **Conditional interpretation**
  - Based on some specific conditions
  - With a set of initial weights, a network is trained to produce an internal representation to be interpreted
- **Individual interpretation**
  - Most methods aim to explain the responses of neural networks only for some specific instances or input patterns
- **Intuitive interpretation**
  - The majority of methods are based on the intuitive and visual interpretation of data such as image data sets, in particular, in the case of CNN

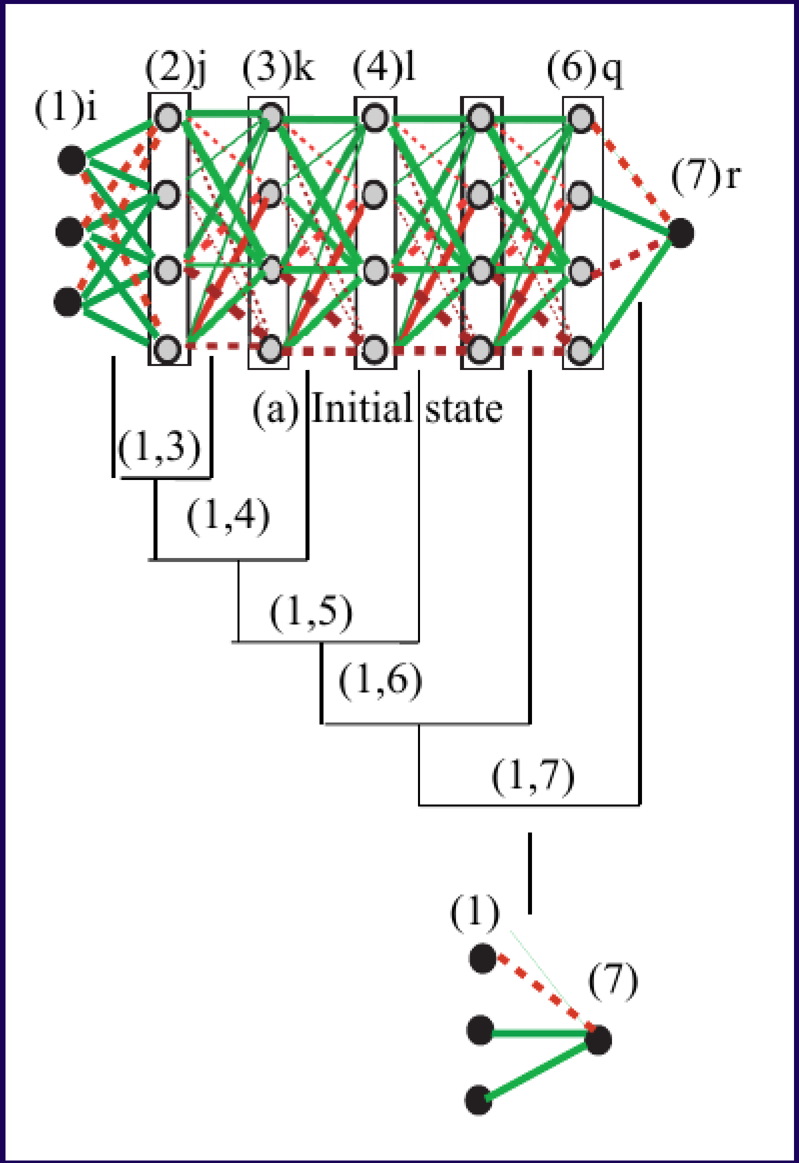
New Interpretation Method: Collective Interpretation

- **Collective and Internal Interpretation**
  - **Network compression**
    - Primarily aims to interpret weights themselves
    - By compressing an original multi-layered neural network into the simplest one without hidden layers
  - **As many internal representations as possible**
    - Individual weights are not dealt with, but all versions of weights are collectively treated
    - By averaging all weights, produced by different initial conditions and different input patterns


Intial state
Different initial conditons
Considering all different representations
Collective interpretation

Compression

- **Collective weights**:
  - All weights in all layers are gradually multiplied and summed to produce collective weights
  - This means that we compute all routes from inputs to outputs
- **No hidden layers**:
  - Any multi-layered neural networks can be reduced to networks without hidden layers
  - The final non-hidden layered networks can be interpreted as the conventional regression analysis



Collective weights

Combination of Individual and Collective Information

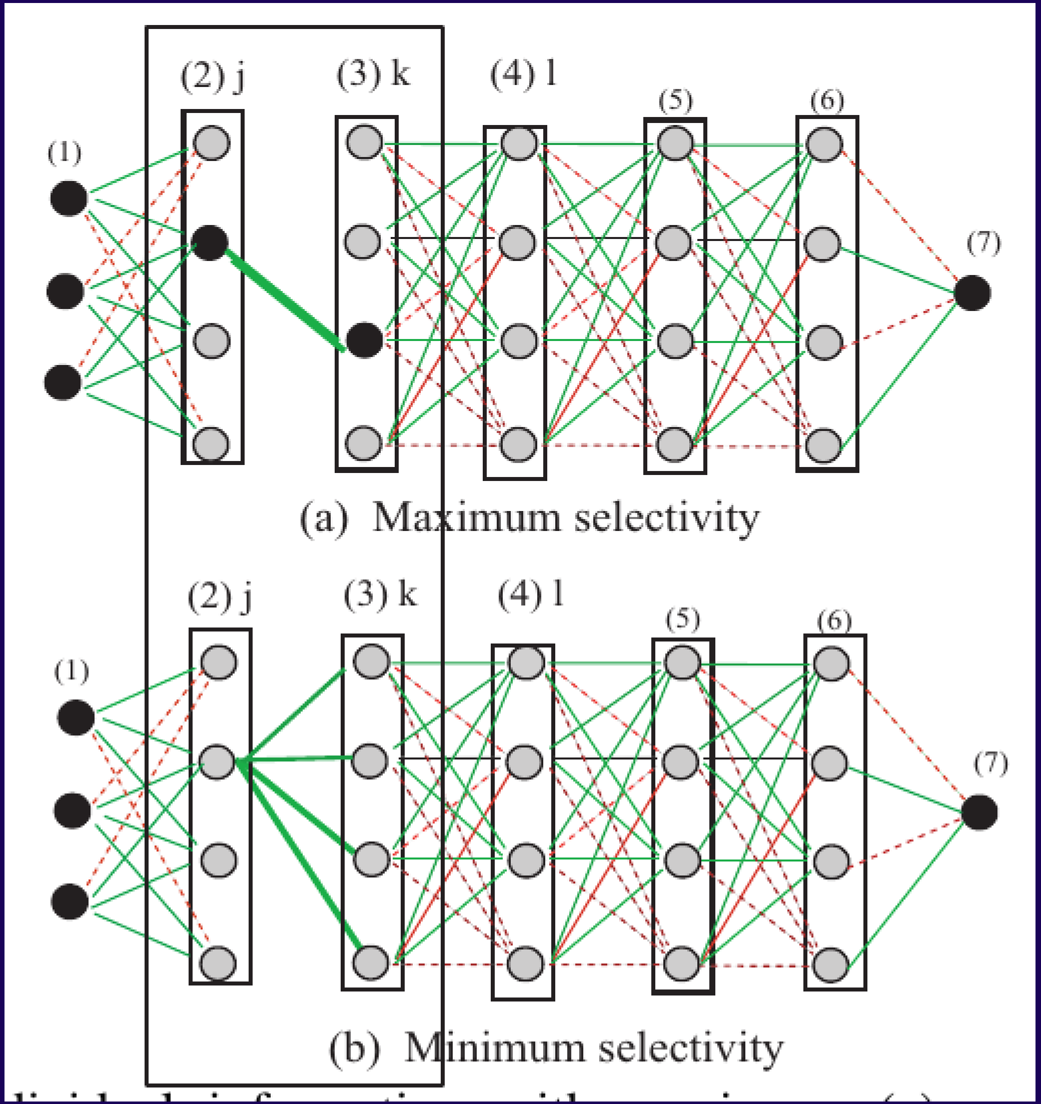
- Flexible Control
  - To control flexibly network configurations

- Individual and Collective Information
  - We introduce individual and collective information

- Ratio of individual to collective
  - By changing the ratio of each information,
  - We can control the final representations very flexiblty

Individual Information Control

- Individual control
  - Neurons and connection weights are individually treated
  - All components are individually and independently changed
  - This control can be used to change internal weights in detail

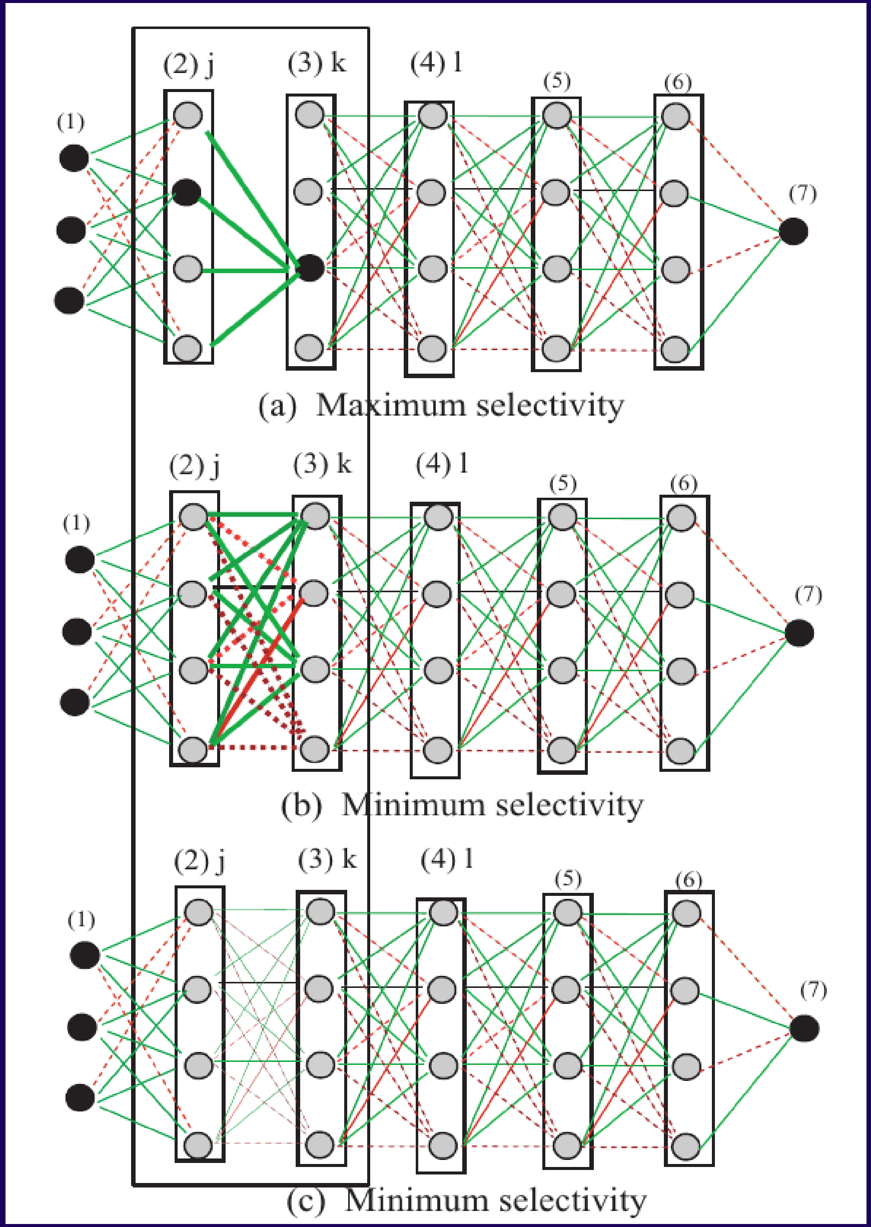


(a) Maximum selectivity

(b) Minimum selectivity

Collective Information Control

- Collective treatment
  - A set of neurons and connection weights are treated as a group
  - Information is defined for this group
  - A group of components are controlled
  - This control can be used to change connection weights roughly



Combination of Individual and Collective Information

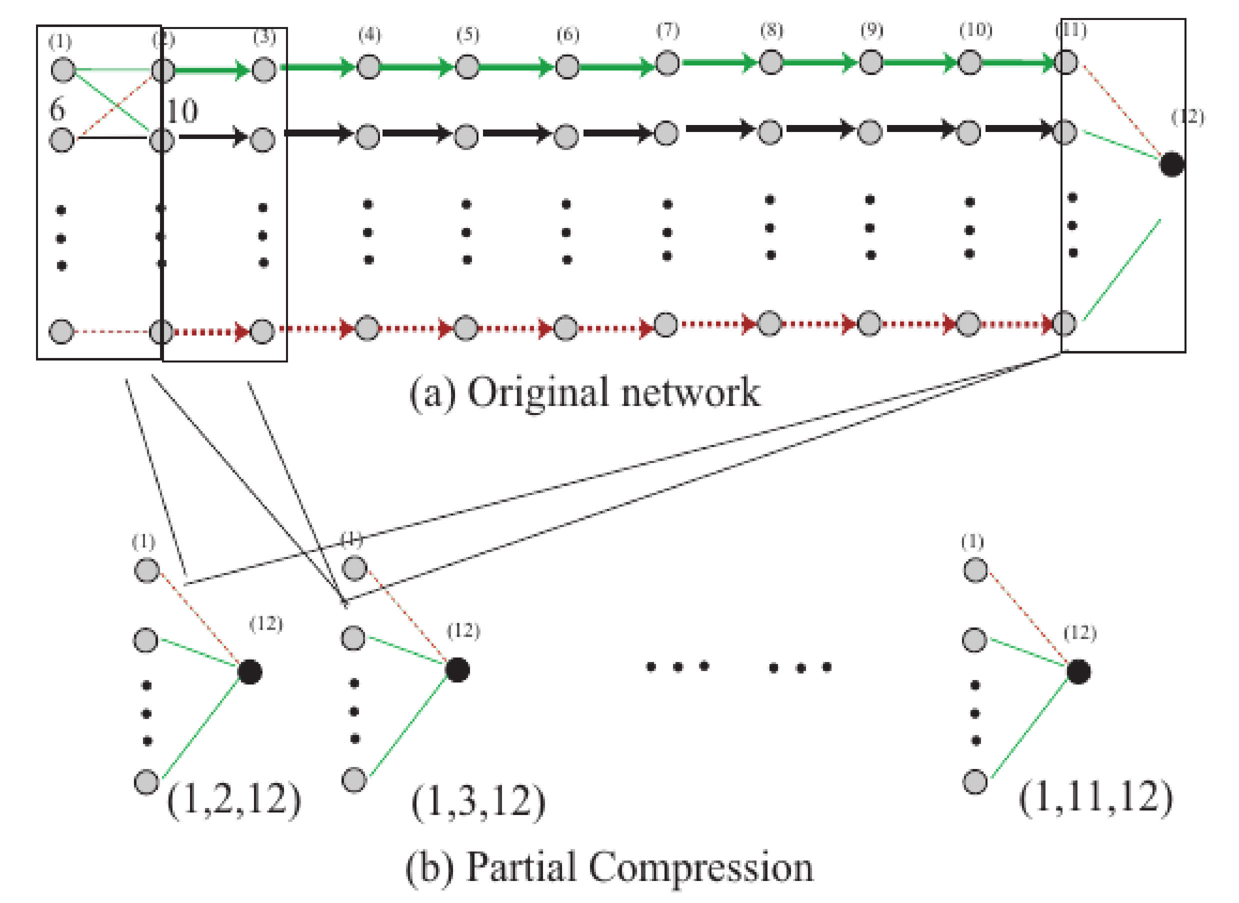
- Controlling the ratio of individual to collective information
  - z (individual) and v (collective) are changed
  - By changing the parameter alpha
  - We can have a variety of internal representations

$$d_{jk}^{(2,3)} = \alpha z_{jk}^{(2,3)} + \bar{\alpha}\bar{v}_{k}^{(3)}$$

Application to Bankruptcy Data Set

- 10 hidden layers with 10 neurons
  - To demonstrate intuitively the performance

- Average results with 10 runs with different initial conditions and different subsets of data set

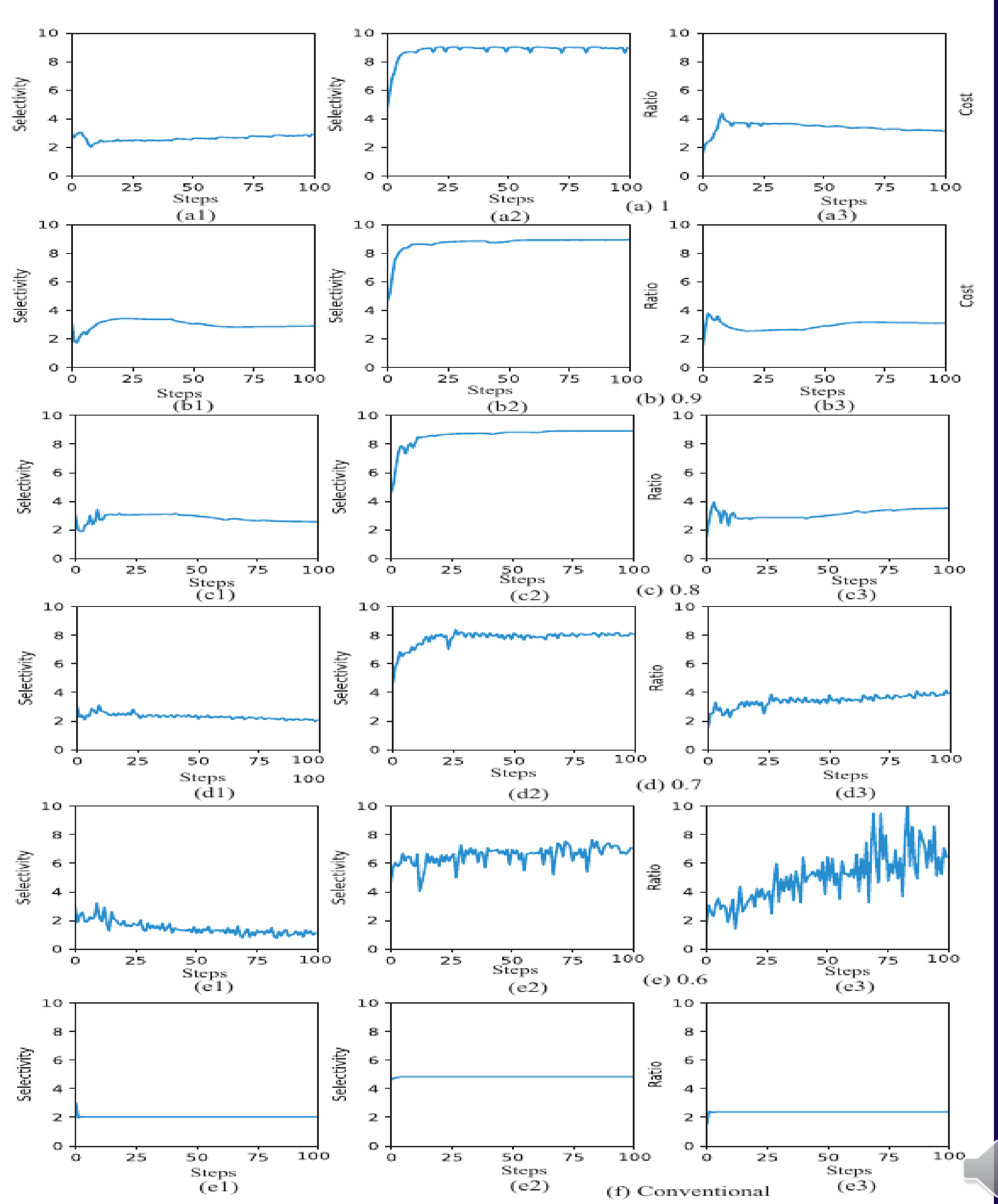


Individual and Collective Information

- Individual information (left)
  - remained to be small
- Collective information (middle)
  - Increased gradually

Cost

- Ratio of individual and collective (right)
  - increased gradually in the end



(a1) (a2) (a) 1 (a3)

(b1) (b2) (b) 0.9 (b3)

(c1) (c2) (c) 0.8 (c3)

(d1) (d2) (d) 0.7 (d3)

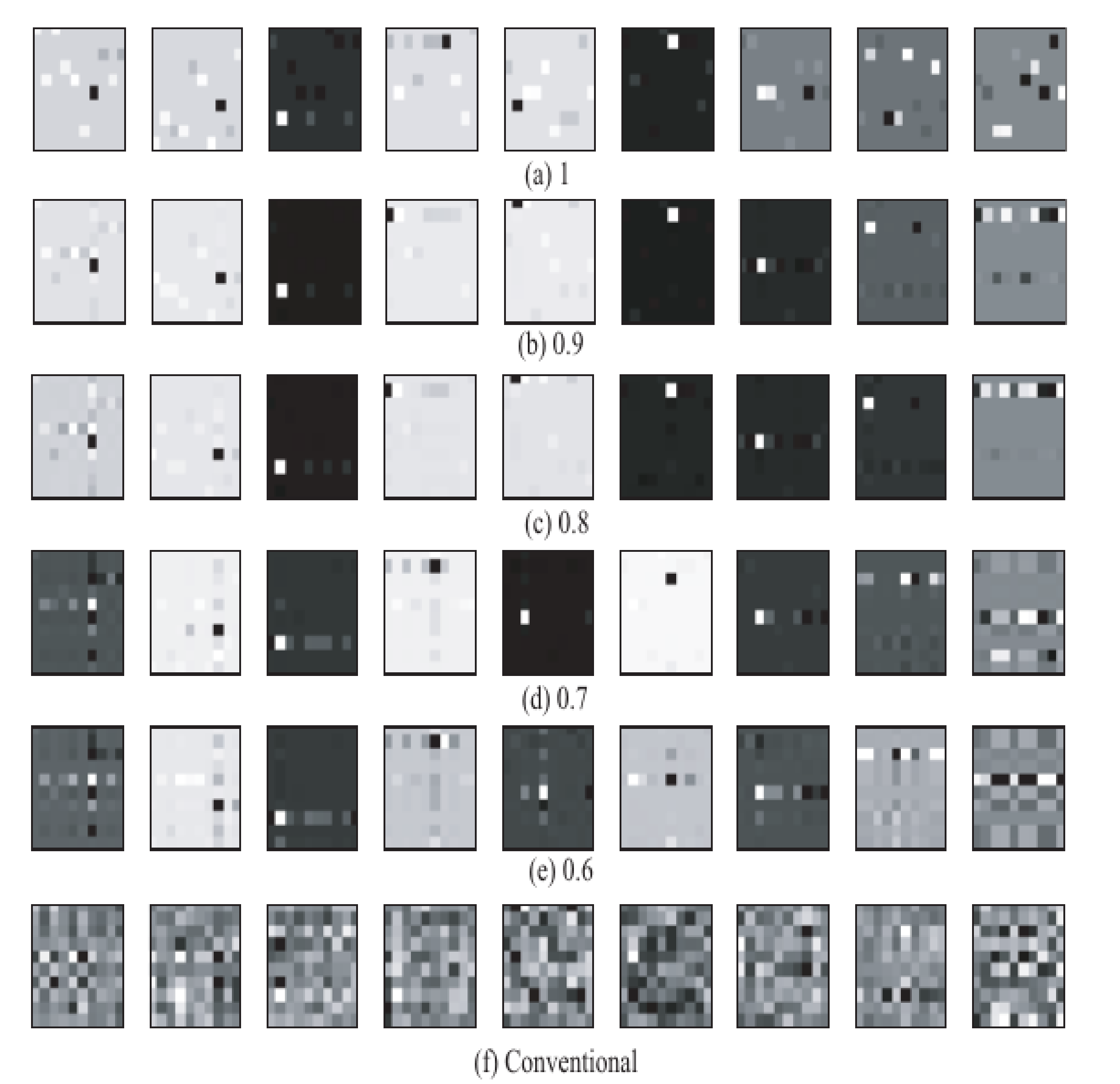
(e1) (e2) (e) 0.6 (e3)

(e1) (e2) (f) Conventional (e3)

Controlling Information

- Sparse weights
  - The number of stronger connection became smaller

- For a limited parameter area
  - Parameter should be from 0.1 to 0.6



(a) 1

(b) 0.9

(c) 0.8

(d) 0.7

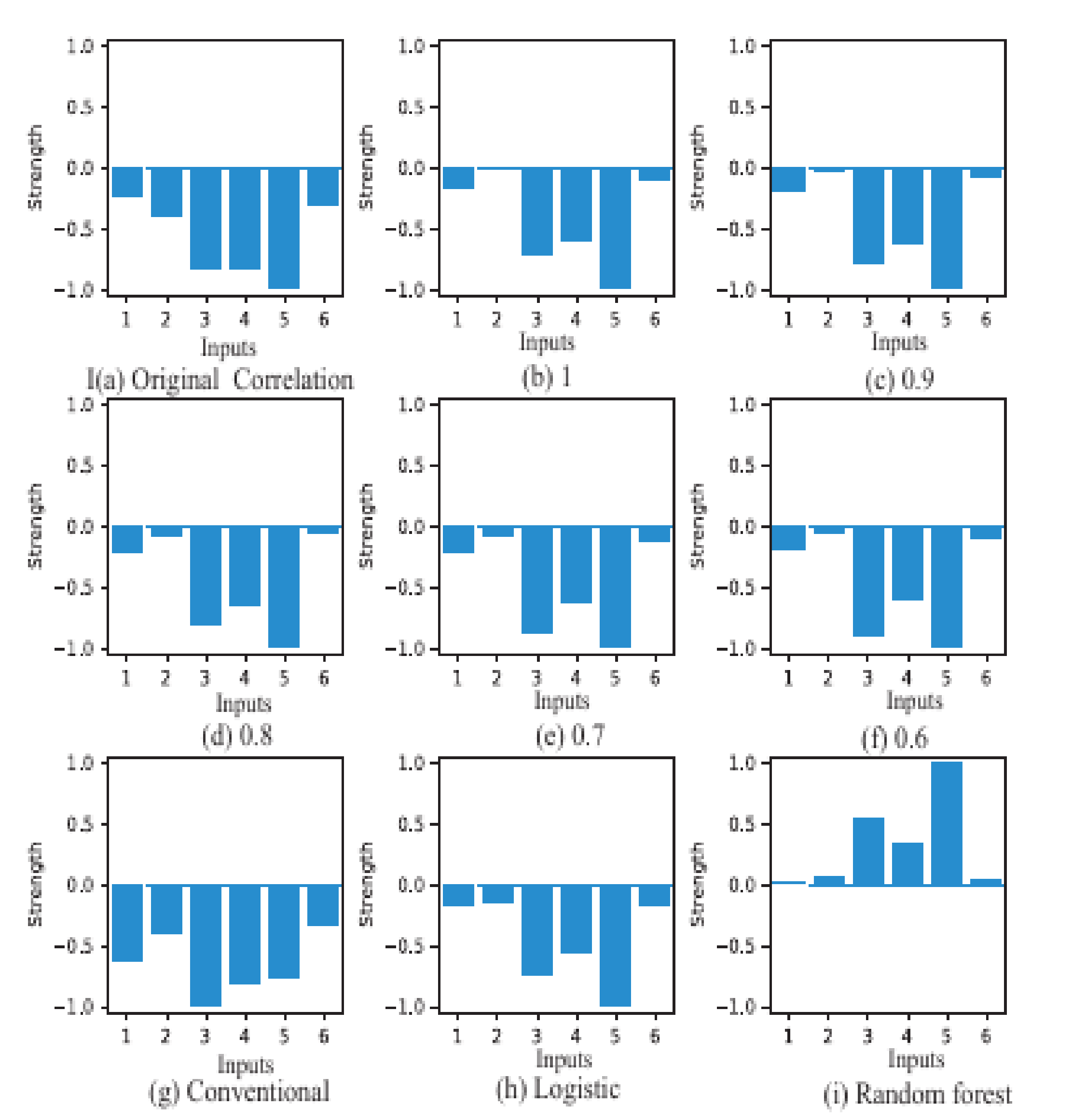
(e) 0.6

(f) Conventional

Collective Weights

- Compressed weights
  - were close to correlation coefficients between inputs and targets of original data set

- Disentangle weights
  - Weights were disentangled to be independently distributed



I(a) Original Correlation

(b) 1

(c) 0.9

(d) 0.8

(e) 0.7

(f) 0.6

(g) Conventional

(h) Logistic

(i) Random forest

Outline

- **Problems of interpretation**
  - Understanding cognitive functions, Critical decision making, improving generalization
  - Difficulty to interpret the main mechanism of neural networks
- **Collective interpretation**
  - Interpreting neural networks, considering all possible representations generated by learning against conditional, individual and intuitive interpretation
- **Network compression**
  - For easy interpretation, we compress multi-layered numeral networks into the simplest ones without hidden layers
  - In addition, partial compression is used to see the state of intermediate layers
- **Combination of Individual and Collective Information**
  - To control information more flexibly for better interpretation, we introduce two types of information, namely, individual and collective information, and control them flexiblty
- **Application to bankruptcy data set**
  - We could detect **linear and independent relations** between inputs and outputs, hidden in complicated non-linear relations

Conclusion

- More exact relations between individual and collective information
- More exact relations between correlation coefficients and collective weights