グラフ構造データを対象とした解釈可能決定集合の拡張

5416063 松山航太

機械学習の解釈性研究

Interpretable Decision Sets

グラフ構造データへの適用

Graph Interpretable Decision Sets（本手法）

H, Lakkaraju, et al : 『Interpretable Decision Sets: A Joint Framework for Description and Prediction』,KDD '16(2016)

1．
ルール集合
の獲得

	○○	××	…	class
D1	No	Yes	…	c1
D2	Yes	No	…	c2
…	…	…	…	…

データセット
D

クラス相関ルール
（if-thenルール）
を用いて
ルールセットを獲得

○○=Y and ××=Y then c1
○○=N then c2
××=N and △△=N then c2
…

ルール集合
S

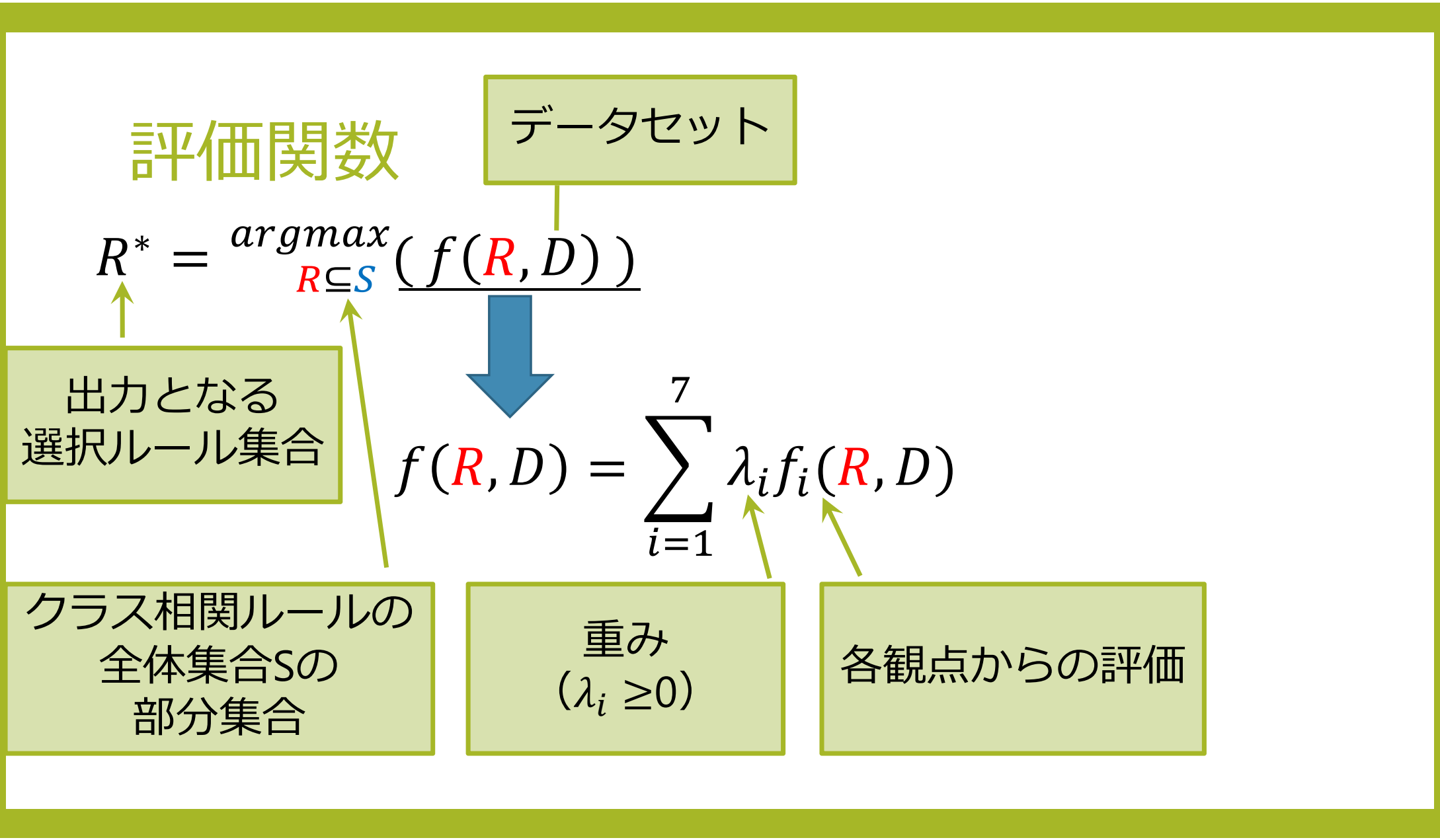
2．
適切な
部分集合の
獲得

○○=Y and ××=Y then c1
××=N and △△=N then c2

解釈可能
決定集合
R

評価関数$f(R, D)$に
従ってRを選択

$$R^* = \underset{R \subseteq S}{argmax}(f(R, D))$$


評価関数
データセット
$R^* = \underset{R \subseteq S}{argmax}\,(f(R, D))$
出力となる
選択ルール集合
$f(R, D) = \sum_{i=1}^{7} \lambda_i f_i(R, D)$
クラス相関ルールの
全体集合Sの
部分集合
重み
($\lambda_i \geq 0$)
各観点からの評価

集合サイズ（小さいほど高評価）

$$f_1(R, D) = |S| - size(R)$$

$$size(R) = 3$$

- ○○=Y and ××=Y then c1
- ××=N and △△=N then c2
- △△=N then c2

ルールの本体部の長さ（短いほど高評価）

$$f_2(R, D) = L_{max} \cdot |S| - \sum_{r \in R} length(r)$$

○○=Y and ××=Y then c1
××=N and △△=N then c2
△△=N then c2

$length(r)$

$$\sum_{r \in R} length(r) = 2 + 2 + 1 = 5$$

ルールの説明範囲の重複度合い（小さいほど高評価）

$$f_3(R,D) = N \cdot |S|^2 - \sum_{\substack{r_i, r_j \in R \\ i \leq j \\ c_i = c_j}} overlap(r_i, r_j) \qquad f_4(R,D) = N \cdot |S|^2 - \sum_{\substack{r_i, r_j \in R \\ i \leq j \\ c_i \neq c_j}} overlap(r_i, r_j)$$

$$overlap(r, r') = cover(r) \cap cover(r')$$

データセットD

	○○	××	…	class
d_1	Y	Y	…	c1
d_2	Y	N	…	c2
d_3	N	Y	…	c2

ルールr
○○=Y then c1

ルールr'
××=Y then c2

cover(r)
d_1
d_2

$overlap(r, r') = 1$
d_1

cover(r')
d_1
d_3

クラスの網羅性（大きいほど高評価）

$$f_5(R, D) = \sum_{c' \in C} 1 \quad (\exists r = (s, c) \in R \; such \; that \; c = c')$$

○

○○=Y and ××=Y then c1
○○=N and △△=Y then c2

c1,c2 両方 =2

×

○○=Y and ××=Y then c1
○○=Y and △△=Y then c1

c1 のみ =1

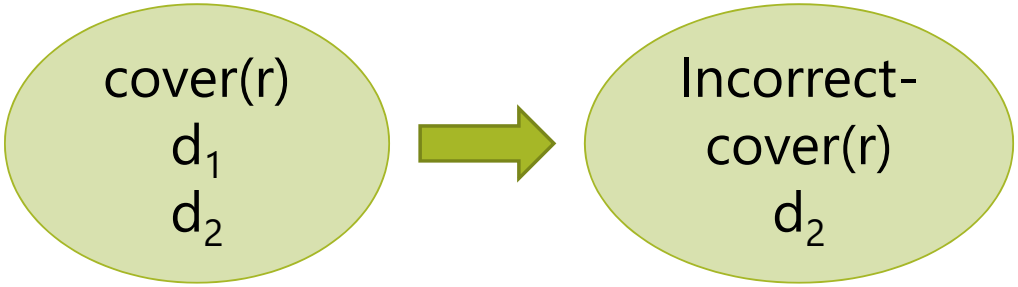
誤分類される事例数（小さいほど高評価）

$$f_6(R, D) = N \cdot |S| - \sum_{r \in R} |incorrect\text{-}cover(r)|$$

	○○	××	…	class
d_1	Y	Y	…	Y
d_2	Y	N	…	N
d_3	N	N	…	N

ルールr

○○=Y then c1



II

1

事例の被覆度（大きいほど高評価）

$$f_7(R, D) = \sum_{(x,y)\in D} 1(|\{r|(x,y) \in correct\text{-}cover(r)\}| \geq 1)$$

データセットD

	○○	××	…	class
d_1	Y	Y	…	c1
d_2	Y	N	…	c2
d_3	N	N	…	c3

r_1:○○=Y then c1
r_2:××=N then c2

correct-cover(r_1) d_1

correct-cover(r_2) d_2 d_3

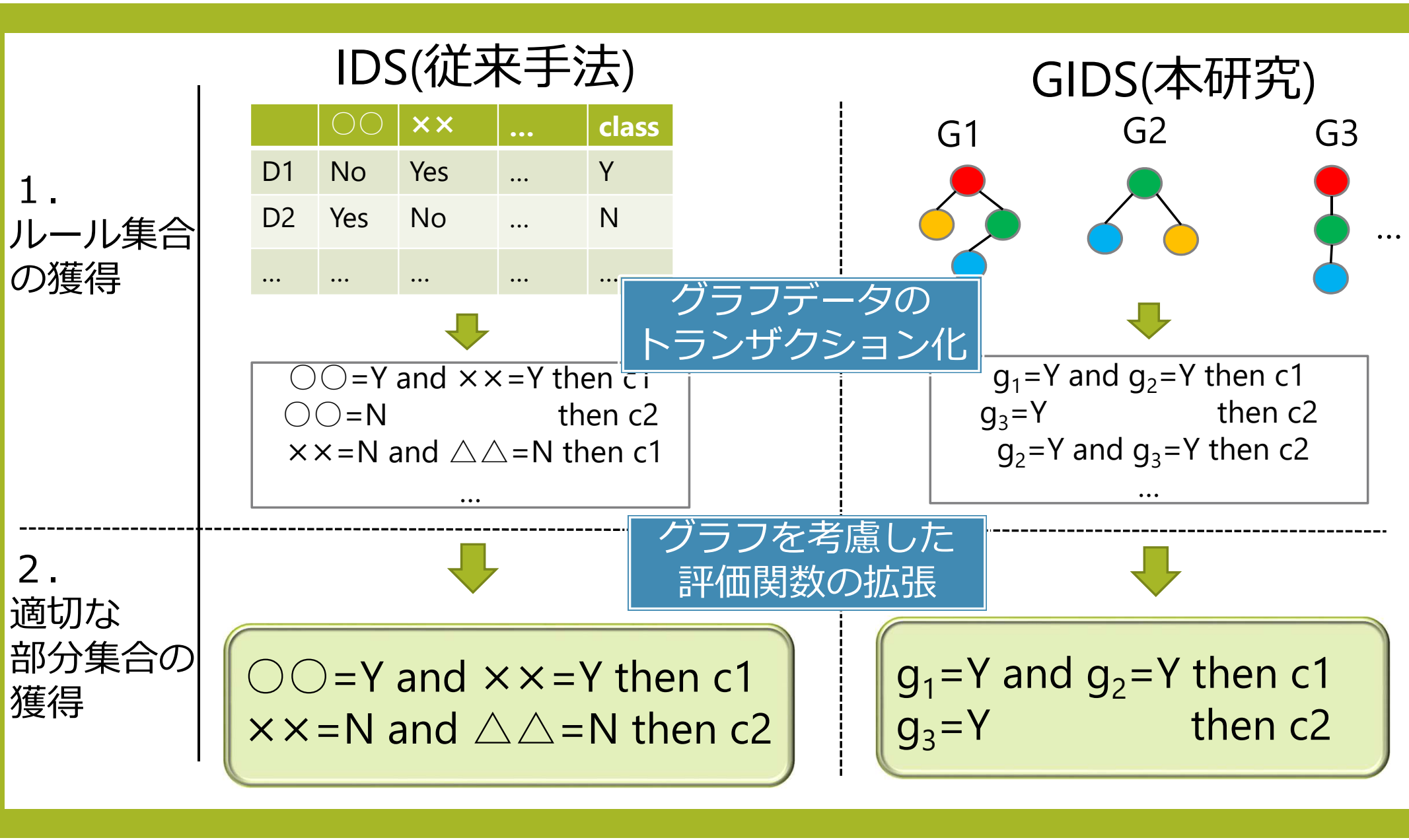
$$f_7(R) = |\{d_1, d_2, d_3\}| = 3$$

$$R^* = \underset{R \subseteq S}{argmax}\, f(R, D)$$

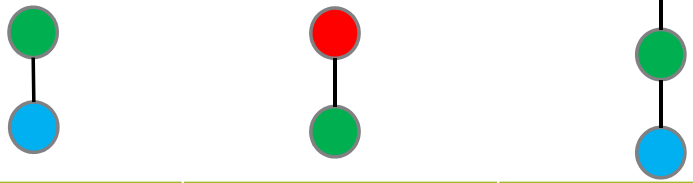
劣モジュラ最大化問題
$f_1(R, D)$から$f_7(R, D)$は全て劣モジュラ関数
既存のアルゴリズム(SLS等)で計算可能

劣モジュラ関数：$f(S \cup \{i\}) - f(S) \geq f(T \cup \{i\}) - f(T)$

台集合:V，　$S, T \subseteq V$，　$S \subseteq T$，　$i \notin \mathrm{Y}$


IDS(従来手法)
GIDS(本研究)
1．
ルール集合
の獲得
○○
××
...
class
D1
No
Yes
...
Y
D2
Yes
No
...
N
...
...
...
...
...
G1
G2
G3
...
グラフデータの
トランザクション化
○○=Y and ××=Y then c1
○○=N then c2
××=N and △△=N then c1
...
g1=Y and g2=Y then c1
g3=Y then c2
g2=Y and g3=Y then c2
...
グラフを考慮した
評価関数の拡張
2．
適切な
部分集合の
獲得
○○=Y and ××=Y then c1
××=N and △△=N then c2
g1=Y and g2=Y then c1
g3=Y then c2

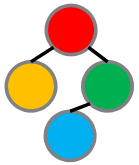
グラフ構造データのトランザクション化



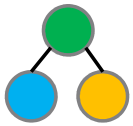
頻出部分グラフ

グラフデータ

G1



G2



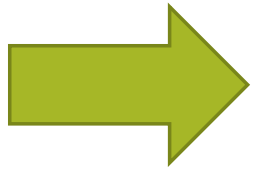
G3

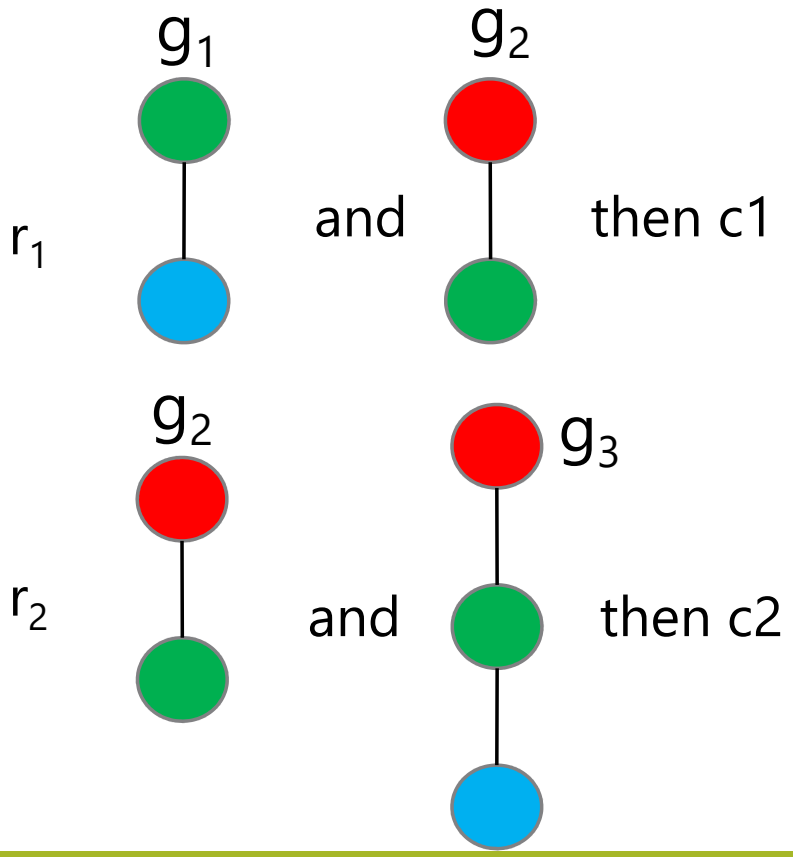
	g_1	g_2	g_3
G_1	Yes	Yes	Yes
G_2	Yes	No	No
G_3	Yes	Yes	Yes

評価関数の拡張

評価関数をルール中の
グラフの形状を考慮するよう拡張

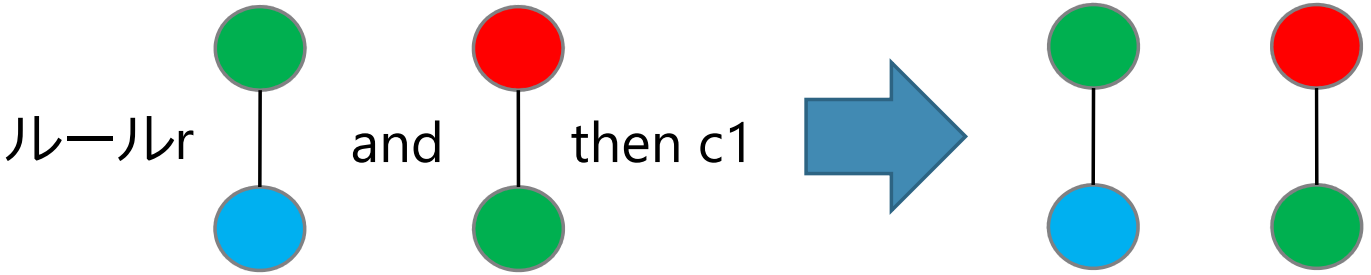
g_1=Y and g_2=Y then c1
g_2=Y and g_3=Y then c2
…





グラフサイズ（小さいほど高評価）

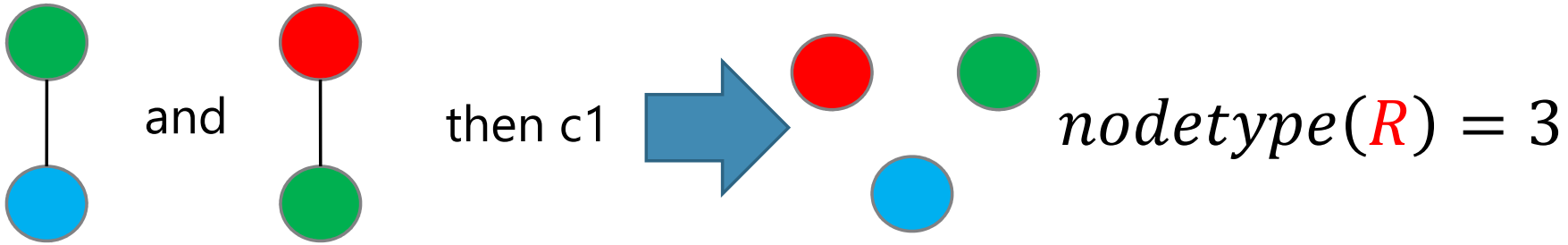
$$f_8(R, D) = \sum_{r \in R} (G_{max} - graphsize(r))$$



$$graphsize(r) = 4 + 2 = 6$$

ノードラベルの種類数（多いほど高評価）

$$f_9(R, D) = nodetype(R)$$



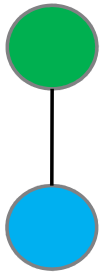
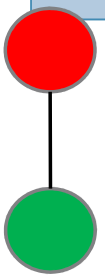
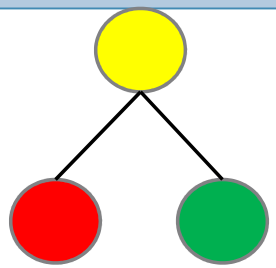
$$nodetype(R) = 3$$

グラフ非類似度（大きいほど高評価）

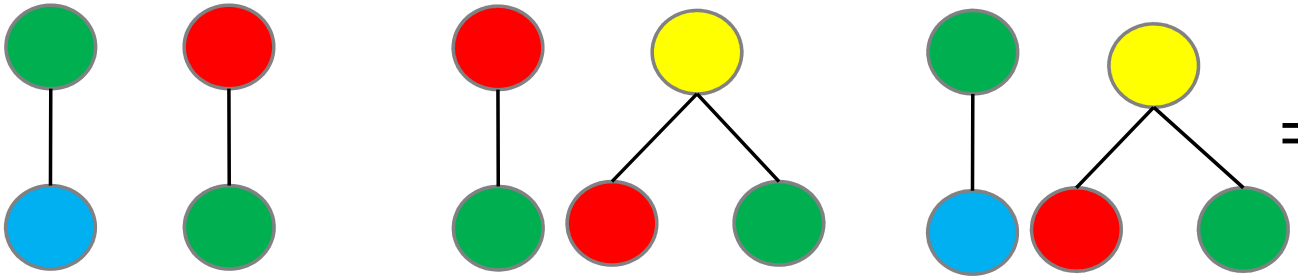
グラフ編集距離
(Graph edit distance)

$$f_{10}(R,D) = \sum_{r \in R} (graph\ dissimilarity(r))$$

$$graph\ dissimilarity(r) = \frac{1}{{}_{|G|}C_2} \sum_{g_i, g_j \in G} GED(g_i, g_j)$$

r =  and  and  then c1

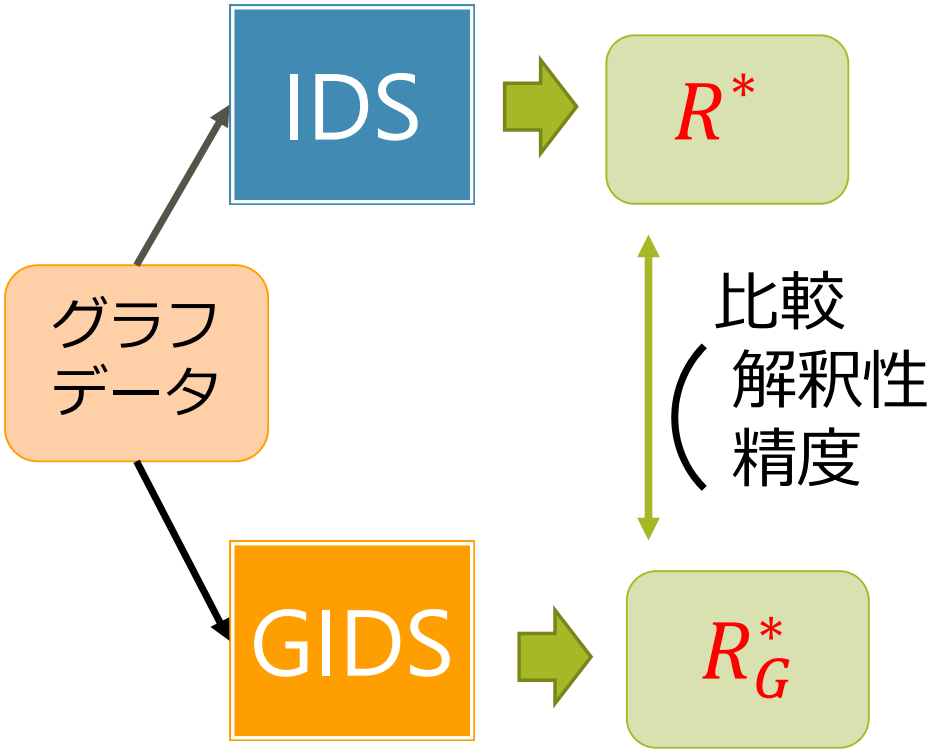
{GED(g_1,g_2) + GED(g_2,g_3) + GED(g_1,g_3) } / 3



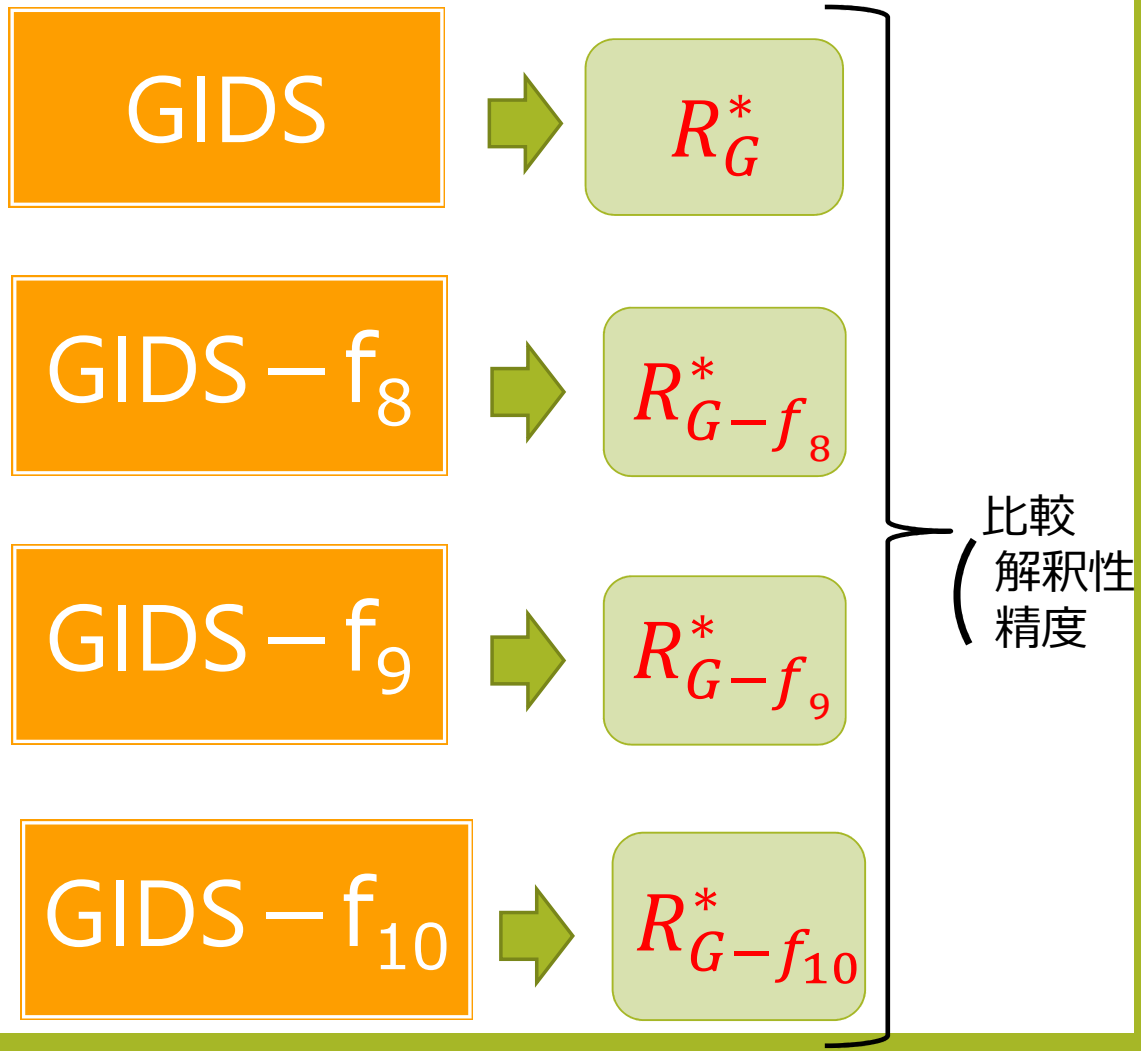
$= graph\ dissimilarity(r)$

実験設定

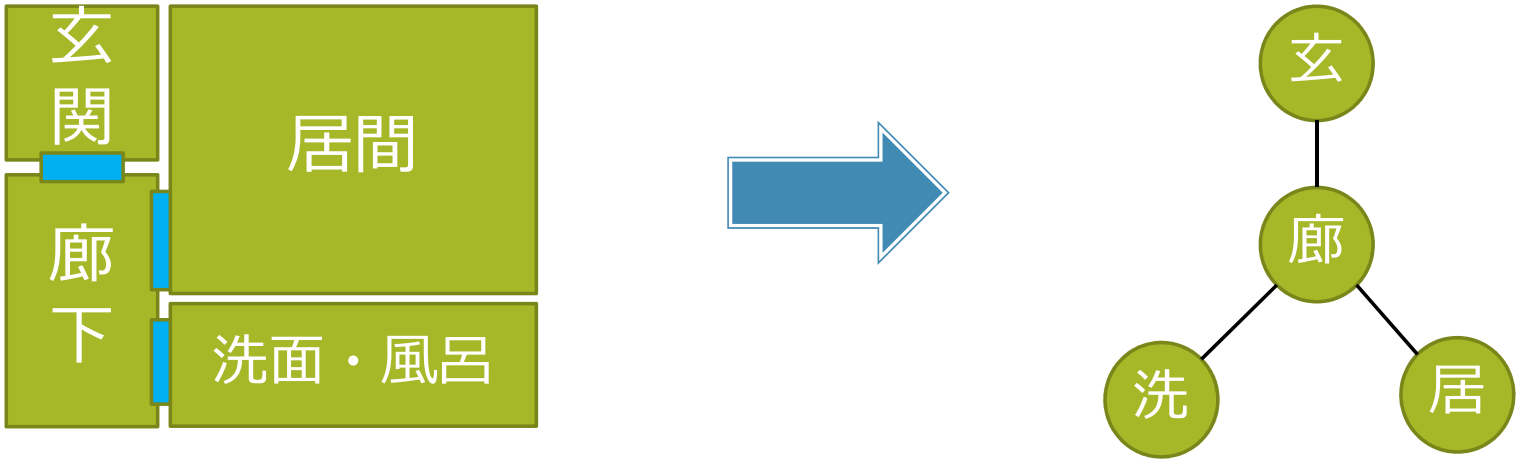
拡張の有無の比較



各関数の有無の比較



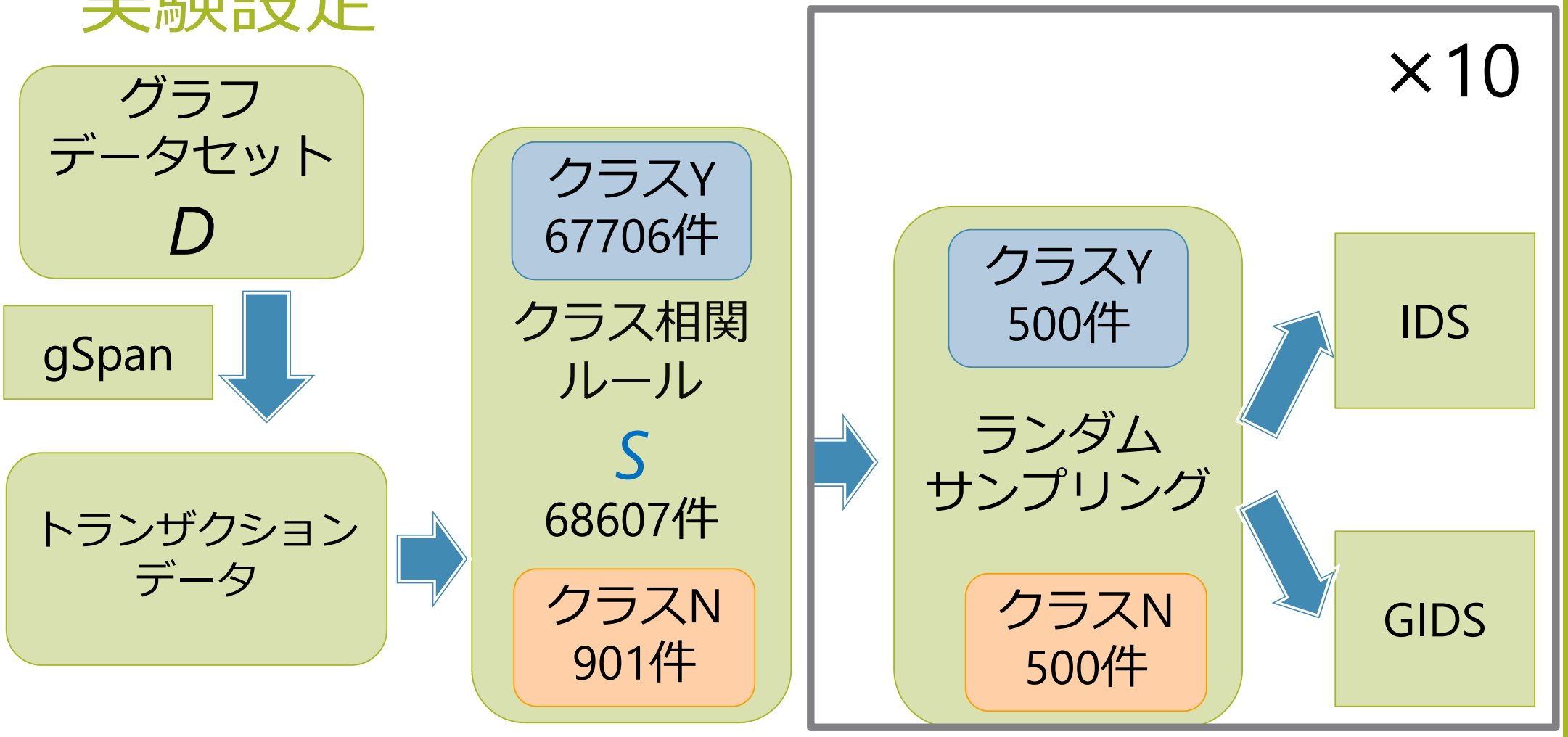
データセット:Homes間取り図グラフ



データ数:480件（Y,N各クラス240件ずつ）

T, Ozaki:Extraction of Characteristic Subgraph Patterns with Support Threshold from Databases of Floor Plans, Cander2019

実験設定


グラフ
データセット
D
gSpan
トランザクション
データ
クラスY
67706件
クラス相関
ルール
S
68607件
クラスN
901件
×10
クラスY
500件
ランダム
サンプリング
クラスN
500件
IDS
GIDS

グラフサイズ

ノード種類数

グラフ非類似度

	IDS	GIDS	¬f_8	¬f_9	¬f_{10}
精度(AUC)	0.52	**0.58**	**0.52**	0.58	0.58
	35	**0.00**	**0.35**	0.00	0.00
	22	**0.11**	**0.22**	0.11	0.11
	95	3.00	**2.95**	3.00	3.00
		4.20	**2.00**	4.20	4.20
	0	1.00	1.00	1.00	1.00
	95	**10.94**	**12.25**	**10.88**	10.94
ノード種数	:f_9 7.80	**9.40**	**7.90**	**9.30**	9.40
非類似度	:f_{10} 0.26	**0.50**	**0.27**	0.50	0.50

・グラフサイズの関数を除くと性能が大きく下がる

・他の関数は誤差レベルの貢献

まとめ

・IDSのグラフ構造データへの拡張
　-精度、解釈性の向上

今後の課題

・追加実験
・λの最適化
・主観評価