DEIM2020 H6-5

ランク付集合ラベルデータの分析範囲と評価基準系列

葛西　正裕†　　古川　哲也††

† 愛知学院大学経済学部　〒 462–8739 愛知県名古屋市北区名城 3-1-1
†† 九州大学大学院経済学研究院　〒 819–0395 福岡県福岡市西区元岡 744
E-mail: †kuzunisi@dpc.agu.ac.jp, ††furukawa@econ.kyushu-u.ac.jp

あらまし　データから有益な情報を得るためにも，データの分析範囲を多面的な視点で特定し分析する必要がある．データの性質などをデータのラベルとして表すモデルでは，概念の階層などに基づいて複数のカテゴリのラベル（集合ラベル）がデータに付される．データに対するラベルの適合度は一般に異なるので，強さを区分するランクをラベルに導入すれば強さを反映した分析が可能になる．しかし，分析対象をラベル集合で与えたとき，ラベル集合とデータのランク付集合ラベルとの関連の強さは様々であり，分析対象に対する分析範囲は明確ではない．本稿は，ラベル集合とデータとの関連の強さに対する評価基準を導出し，評価基準を満たすデータを分析範囲にすることで，分析対象と分析範囲の関係を明確にする．さらに，評価基準の強さの順序を明らかにすることで，強さの比較が可能な評価基準の集合を系列として提案する．評価基準系列により比較可能な分析範囲の候補を段階的かつ一元的に提示できるので，ランク付集合ラベルデータを詳細に分析できるようになる．

キーワード　ランク付集合ラベルデータ，ランク，マルチラベル，分析範囲，評価基準，データ分析，データモデル

1　はじめに

収集したデータから有益な情報を得ることの重要性が増している．データの分析範囲によって分析結果が異なる場合があり，新しい知見を得られることがある．データから有益な情報を得るためにも，分析範囲を一面的ではなく多面的な視点で特定し，分析結果を適切に検証しなければならない．

収集されるデータの種類は数値データ，テキスト，画像，音声，動画など多岐に及び，多様なデータに対して分析が行われている [10] [11] [12] [13]．データを分析するにはデータを適切に構成しておく必要があり [3] [7]，データの性質をデータのラベル（タグ）で表すモデルでは，データには分類されるカテゴリ（クラス）のラベルが付されている [1] [5] [6] [9]．実世界の多くの領域では，カテゴリが示す概念の関係は木構造や DAG で表現でき [8]，ラベル間に階層的な関係がある場合が多い [1] [5] [6]．

ラベルはデータの属性値に該当するので，分析対象をラベルで与えてデータのラベルとの関係で分析範囲が特定される．階層的な関係があるラベルでは，分析対象を表すラベルに対し，同位もしくは下位概念のラベルのデータが分析範囲になる．例えば，企業で構成されるデータにおいて，展開地域や業種といった属性の値がラベルで付されていれば，分析対象を “日本” というラベルで与えたとき，日本に属する地域である “愛知” や “福岡” といったラベルのデータが分析範囲に含まれる．

データのラベルが単数であれば，分析対象に対する分析範囲は明確である．しかし，企業データでいえばグローバル経済の進展とともに複数の地域で展開している企業が多いように，データはただ 1 つのカテゴリに分類できる単純なものばかりではない．分類される複数のカテゴリのラベルがデータに付されており [1] [5] [6] [9]，データのラベルは集合ラベルになる．分析対象をラベル集合で与えたとき，どのようなデータを分析範囲に含めるのかという分析対象の解釈には様々な考え方があり，同じ分析対象であってもその解釈に応じて複数の分析範囲が生じる [2]．さらに，データに対するラベルの適合度は一般に異なる．企業が複数の地域に展開している状況を考えると，本社や研究開発拠点があるような関連が強い地域もあれば，販売代理店を配置している程度の関連が弱い地域もある．強さを区分するランクをデータのラベルに導入すれば強さを反映した分析が可能になるが，分析対象の解釈はより複雑になり，分析対象と分析範囲の関係が不明瞭のままでは的確な分析ができない．

分析対象を表すラベル集合とデータのランク付集合ラベルの関係は様々なので，ラベル集合とデータとの関連の強さは同じではない．例えば, 日本や中国に関連する企業を分析する際，両国に強く関連するデータ以外に，両国のどちらか一方のみに強く関連するデータ，両国に強く関連するが米国にも強く関連するデータなどがあり，ラベル集合 $\mathcal{L} = \{$ 日本, 中国 $\}$ で分析範囲に含まれるデータを特定するとき，$\mathcal{L}$ への関連の強さは様々である．ラベル集合とデータとの関連の強さを評価する基準があれば，関連の強さに基づき分析範囲を特定できる．本稿は，ラベル集合とデータとの関連の強さに対する評価基準を述べる．

評価基準は分析対象の解釈に対応し，評価基準を満たすデータを分析範囲にすることで，分析対象と分析範囲の関係は明確になる．分析範囲は評価基準の条件で説明できるので，関連の強さに基づいた分析が可能になる．例えば，日本と中国の経済的なつながりに関心があり，$\mathcal{L}$ を日本と中国の両国に関連するデータと解釈すれば，$\mathcal{L}$ のすべてのラベルに関連するという条件の評価基準 LA を満たすデータが分析範囲になる．LA を満たす企業群の営業利益の平均値が全企業と比較して高ければ，

両国へ展開している企業の方が営業利益が高いと推測できる．

データのラベルは関連が強いラベルと弱いラベルからなる．評価基準 LA を満たすデータ d_i（$i = \{1,2,3\}$）のラベルを関連が強いラベルの集合と関連が弱いラベルの集合で表し，d_1 : { 東京, 北京 }, ϕ，d_2 : { 東京 }, { 北京 }，d_3 : {$N.Y.$}, { 東京, 北京 } とする．d_1，d_2，d_3 の順で $\mathcal{L}$ との関連が強い．$\mathcal{L}$ のすべてのラベルに強く関連するという条件の評価基準 PA を用いて，PA を満たす d_1 のような企業群がさらに高い営業利益であれば，$\mathcal{L}$ への関連がより強い企業ほど利益を上げていると推測できる．評価基準 LA と PA による営業利益の平均値の差は，評価基準の条件の差，すなわち強く関連するという条件が加わったことによるものであり，日本と中国へ重点的に展開していることの重要性が示唆される．

評価基準 LA と PA のように強さの比較が可能な評価基準を用いて分析範囲を特定し分析結果を比較すれば，評価基準の条件の差，すなわち関連の強さの差で説明できるので分析結果を適切に検証できる．本稿は，評価基準の強さの順序を明らかにすることで，強さの比較が可能な評価基準の集合を系列として提案する．評価基準系列を用いることで比較可能な分析範囲の候補を段階的かつ一元的に提示できる．系列を用いることで多面的な視点で分析範囲を特定したうえで分析結果を適切に検証できるので，ランク付集合ラベルデータを詳細に分析できる．

本稿は以下のように構成される．2 節では，データに対するラベルの適合度を表すランクを導入し，データとラベル集合との関連の強さに対する評価基準を議論する．3 節では，評価基準の強さの関係を明らかにし，4 節では，それをもとに評価基準の積で得られる新たな評価基準を導出する．5 節では，評価基準の強さの順序を明らかにすることで，強さの比較が可能な評価基準の集合を系列として定義し，関連の強さを段階的かつ一元的に評価できる理論的枠組みを与える．6 節では，系列を用いたデータ分析について述べる．7 節はまとめである．

2 ラベル集合とオブジェクトとの関係

分析に供されるデータには，概念の階層などに基づいてラベル集合が付されている．ラベル集合の要素間には属性ごとに階層的な関係があることが多く [1] [5] [6]，例えば，属性が業種であれば，製造業，輸送機器，四輪自動車というように，地域であれば，日本，中部，愛知というように上下関係がある．本稿では，議論を簡単にするために，属性は単一とし，データには最下層のラベルが付されているものとする．

1 件のデータをオブジェクトを o，ラベルを l，ラベル集合を $L = \{l_1, l_2, \cdots, l_n\}$ とする．o に付されたラベル集合を集合ラベルと呼び $L(o)$ で表す．ラベル l_1，l_2 に対し，l_1 が l_2 の下位概念のラベルならば，l_1 は l_2 の下位（l_2 は l_1 の上位）であり，$l_1 \prec l_2$ で表す．l_1 が l_2 の下位または l_1 と l_2 が等しい（$l_1 \preceq l_2$）とき，l_1 は l_2 に関連するという．また，ラベル l がラベル集合 L 中のいずれかのラベルに関連するとき，l は L に関連するという．

ラベル集合 L，L' に対し，L' に関連する L 中のラベルの集合を $Ring_{L'}(L) = \{l \mid l \in L, \exists l' \in L', l \preceq l'\}$，$L$ のラベルが関連する L' 中のラベルの集合を $Red_L(L') = \{l' \mid l' \in L', \exists l \in L, l \preceq l'\}$ で表す．$Ring_{L'}(L) \neq \phi$ は $Red_L(L') \neq \phi$ と等価である．

オブジェクトとの関係を議論する際に用いるラベル集合を $\mathcal{L}$（$\mathcal{L} \neq \phi$）で表す．オブジェクト o の $L(o)$ のラベルがラベル l に関連するとき，o は l に関連するといい，$L(o)$ のラベルがラベル集合 $\mathcal{L}$ に関連するとき，すなわち $Ring_{\mathcal{L}}(L(o)) \neq \phi$ のとき，o は $\mathcal{L}$ に関連するという．また，$\mathcal{L}$ に関連するオブジェクトの集合を $\overline{\mathcal{L}}$ で表す．

ラベル集合 $\mathcal{L}$ に対し，$\mathcal{L}$ に強く関連するオブジェクトもあれば関連が弱いオブジェクトもあるので，$\mathcal{L}$ とオブジェクトとの関連の強さは同じではない．$\mathcal{L}$ とオブジェクトとの関連の強さを評価する基準があれば，関連の強さに基づいて分析範囲を特定できる．オブジェクト o_1，o_2 に対し，o_2 の方が o_1 よりもラベル集合 $\mathcal{L}$ との関連が強いことを $o_1 <_{\mathcal{L}} o_2$，$\mathcal{L}$ が明確であれば $o_1 < o_2$ で表す．

［定義 1］ オブジェクト o_1，o_2 と条件 cnd に対し，o_2 は cnd を満たし o_1 は cnd を満たさないとき $o_1 <_{\mathcal{L}} o_2$ ならば，cnd はラベル集合 $\mathcal{L}$ とオブジェクトとの関連の強さの評価基準である．

ラベル集合 $\mathcal{L}$ とオブジェクト o との関連の強さは，$\mathcal{L}$ と $L(o)$ との関係で考えることができる．関係には 4 種類が考えられ，$\mathcal{L}$ と $L(o)$ のどちらに着目するのか，いずれかのラベルとすべてのラベルのどちらで考えるかによるものである．$\mathcal{L}$ のいずれかのラベルに対し $L(o)$ のラベルが関連する関係（$Red_{L(o)}(\mathcal{L}) \neq \phi$）は，$L(o)$ のいずれかのラベルが $\mathcal{L}$ に関連する関係（$Ring_{\mathcal{L}}(L(o)) \neq \phi$）と同じ関係である．また，$\mathcal{L}$ のすべてのラベルに対し $L(o)$ のラベルが関連する関係（$Red_{L(o)}(\mathcal{L}) = \mathcal{L}$）と $L(o)$ のすべてのラベルが $\mathcal{L}$ に関連する関係（$Ring_{\mathcal{L}}(L(o)) = L(o)$）がある．$L(o)$ のすべてのラベルが $\mathcal{L}$ に関連するということは，$L(o)$ には $\mathcal{L}$ と無関係なラベル l（$l \not\preceq l', l' \not\preceq l, \forall l' \in \mathcal{L}$）に関連するラベルは存在しないと言い換えられる．これらの関係から $\mathcal{L}$ と $L(o)$ との関連の強さは以下のように考えることができる．ラベル集合 $\mathcal{L}$ とオブジェクト o_1，o_2 に対し，

（1） o_2 は $\mathcal{L}$ に関連するが o_1 は関連しない，

（2） o_2 は $\mathcal{L}$ のすべてのラベルに関連するが o_1 は $\mathcal{L}$ のいずれかのラベルに関連しない，

（3） o_2 は $\mathcal{L}$ と無関係なラベルに関連しないが o_1 は無関係なラベルに関連する，

のいずれかを満たせば o_2 は o_1 よりも $\mathcal{L}$ に強く関連する．

［例 1］ 図 1 は，ラベル集合 $\mathcal{L}_1$ = { 中部, 九州 } とオブジェクトとの関係の例を示している．オブジェクト o_1 は，$L(o_1)$ の “東京” が $\mathcal{L}_1$ の “中部” と “九州” のどちらにも関連しないので $\mathcal{L}_1$ に関連しない．一方，オブジェクト o_2 は，$L(o_2)$ の “福岡” が $\mathcal{L}_1$ の “九州” に関連するので $\mathcal{L}_1$ に関連する．よって，o_2 は o_1 よりも $\mathcal{L}$ に強く関連する．オブジェクト o_3 は，$L(o_3)$ の “愛知” が $\mathcal{L}_1$ の “中部” に関連するので $\mathcal{L}_1$ に関連するが，“九州” には関連しない．一方，オブジェクト o_4 は，$L(o_4)$ の “愛知” と “福岡” がそれぞれ $\mathcal{L}_1$ の “中部” と “九州” に関連する，す

なわち $\mathcal{L}_1$ のすべてのラベルに関連するので，o_4 は o_3 よりも $\mathcal{L}_1$ に強く関連する．オブジェクト o_5 は，$L(o_5)$ の “東京” が “中部” にも “九州” にも関連しないので $\mathcal{L}_1$ と無関係なラベルに関連する．一方，オブジェクト o_6 は，$L(o_6)$ のすべてのラベルである “愛知” と “岐阜” が “中部” に関連し $\mathcal{L}_1$ と無関係なラベルに関連しないので，o_6 は o_5 よりも $\mathcal{L}_1$ に強く関連する．

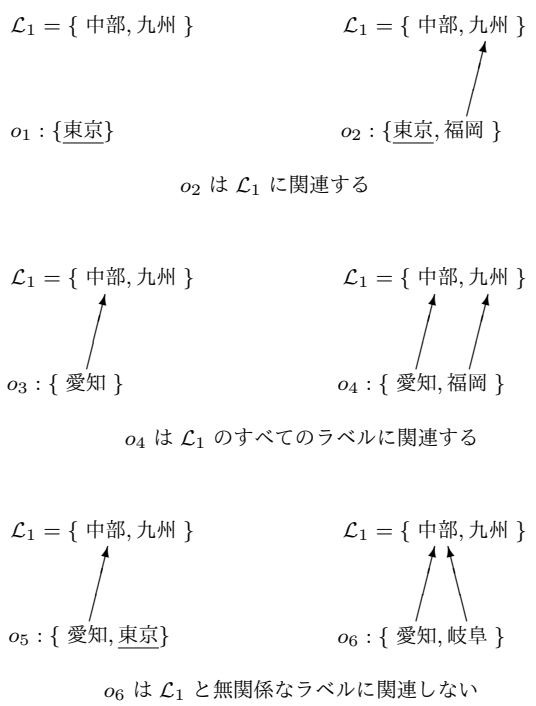


図 1　ラベル集合とオブジェクトとの関連の強さ

このようなラベル集合とオブジェクトの関係は，次のようなラベル集合とオブジェクトとの関連の強さに対する評価基準になる．ラベル集合 $\mathcal{L}$ とオブジェクト o に対し，

LE：　o は $\mathcal{L}$ に関連する（$Ring_{\mathcal{L}}(L(o)) \neq \phi$），

LA：　o は $\mathcal{L}$ のすべてのラベルに関連する（$Red_{L(o)}(\mathcal{L}) = \mathcal{L}$），

LN：　o は $\mathcal{L}$ と無関係なラベルに関連しない（$Ring_{\mathcal{L}}(L(o)) = L(o)$）

は $\mathcal{L}$ と o との関連の強さに対する評価基準である．

オブジェクト o に対する $L(o_i)$ のラベルの適合度は一般に異なるので，$L(o_i)$ の各ラベルに o との関連の強さを表すランクを与える．強さの区分が多いほど関連の強さを精緻に分析に反映できるが，本稿では議論を簡単にするため 2 つの区分のランクを考え，o に強く関連する $L(o)$ 中のラベルを主ラベル（Primary Label），主ラベルほどではないが o に関連する $L(o)$ 中のラベルを副ラベル（Secondary Label）とする．

ランク付集合ラベルのうち主ラベルの集合を $P(o)$，副ラベルの集合を $S(o)$ で表す．ラベルにランクが与えられた集合ラベルをランク付集合ラベルといい，これより，オブジェクトの集合ラベルはランク付集合ラベルとする．オブジェクト o のランク付集合ラベル $L(o)$ の各ラベルは，主ラベルもしくは副ラベルのいずれかなので，$L(o) = P(o) \cup S(o)$ かつ $P(o) \cap S(o) = \phi$ である．また，o の $L(o)$ 中のラベルで最も強く関連しているラベルは主ラベルなので，主ラベルは少なくとも 1 つは存在し，$P(o) \neq \phi$ である．

評価基準 LE，LA，LN は，$L(o)$ がランク付集合ラベルであっても評価基準である．これらの評価基準は $\mathcal{L}$ と $L(o)$ の関係であるが，$\mathcal{L}$ と $P(o)$ との関係からも同様の評価基準を得ることができる．オブジェクト o とラベル集合 $\mathcal{L}$ に対し，

PE：　o は $\mathcal{L}$ に強く関連する（$Ring_{\mathcal{L}}(P(o)) \neq \phi$），

PA：　o は $\mathcal{L}$ のすべてのラベルに強く関連する（$Red_{P(o)}(\mathcal{L}) = \mathcal{L}$），

PN：　o は $\mathcal{L}$ と無関係なラベルに強くは関連しない（$Ring_{\mathcal{L}}(P(o)) = P(o)$）

は o と $\mathcal{L}$ との関連の強さに対する評価基準である．

ラベル集合 $\mathcal{L}$ とオブジェクト o との関連の強さに関して，主ラベルは副ラベルよりも優先されるので，$\mathcal{L}$ と $S(o)$ との関係に基づく条件は評価基準にならない．例えば，$\mathcal{L}_2 =$ { 中部 } に対し，$P(o_7) =$ { 愛知 }，$S(o_7) =$ { 福岡 } であるようなオブジェクトを o_7，$P(o_8) =$ { 福岡 }，$S(o_8) =$ { 愛知 } であるようなオブジェクトを o_8 とする．o_8 は $Ring_{\mathcal{L}}(S(o_8)) \neq \phi$ を満たし o_7 はその条件を満たさないが，PE を満たす o_7 は $\mathcal{L}$ に強く関連し，PE を満たさない o_8 は $\mathcal{L}$ に強く関連しないので $o_8 < o_7$ である．同様に，$Red_{S(o)}(\mathcal{L}) = \mathcal{L}$ という条件と $Ring_{\mathcal{L}}(S(o)) = S(o)$ という条件も評価基準にならない．よって，$\mathcal{L}$ と $S(o)$ との関係に基づく条件は評価基準にならない．

3　評価基準の強さの関係

評価基準が他の評価基準を含意する場合がある．含意関係にある評価基準には強さの関係があると考えることができる [4]．本節は，評価基準の含意関係を検討することで，評価基準の強さの関係を明らかにする．

オブジェクト o_1 と o_2 のラベル集合 $\mathcal{L}$ に対する関連の強さが，評価基準 cnd_1 を満たすかどうかよりも評価基準 cnd_2 を満たすかどうかで決まるならば，cnd_2 は cnd_1 よりも強い $\mathcal{L}$ の評価基準であるとする．すなわち，o_1 や o_2 が cnd_1 を満たすかどうかに関わらず o_2 が cnd_2 を満たし o_1 が cnd_2 を満たさないとき $o_1 < o_2$ ならば，cnd_2 は cnd_1 よりも強い $\mathcal{L}$ の評価基準である．これは，o_1 が cnd_1 を満たしていても, すなわち，o_1 が cnd_1 の評価基準で o_2 よりも強い関連がある可能性があったとしも，cnd_2 を満たさないことで関連の強さが決まることによる．

［定義 2］　ラベル集合 $\mathcal{L}$ と評価基準 cnd_1，cnd_2 に対し，オブジェクト o_2 は cnd_2 を満たし，cnd_1 を満たすオブジェクト o_1 が cnd_2 を満たさないとき $o_1 <_{\mathcal{L}} o_2$ であるならば，cnd_2 は cnd_1 よりも強い $\mathcal{L}$ の評価基準であり，cnd_1 と cnd_2 に強さの関係があるといい，$cnd_1 <_{\mathcal{L}} cnd_2$ で，$\mathcal{L}$ が明確なときは $cnd_1 < cnd_2$ で表す．

ラベル集合 $\mathcal{L}$ に関連するオブジェクト集合 $\overline{\mathcal{L}}$ で評価基準 cnd を満たすオブジェクトの集合を $\overline{\mathcal{L}}^{cnd}$(=$\{o \mid o \in \overline{\mathcal{L}}, o$ は cnd を満たす $\}$) で表す．$\overline{\mathcal{L}}$ は $\mathcal{L}$ に関連するオブジェクトの集合なので，$\overline{\mathcal{L}} = \overline{\mathcal{L}}^{LE}$ である．

評価基準の強さの関係は，評価基準を満たすオブジェクト集

合の包含関係で判断できる．

［補題 1］ ラベル集合 $\mathcal{L}$ と評価基準 cnd_1，cnd_2 に対し，$\overline{\mathcal{L}}^{cnd_2} \subseteq \overline{\mathcal{L}}^{cnd_1}$ であることは，$cnd_1 < cnd_2$ であることの必要十分条件である．

（証明） $\overline{\mathcal{L}}^{cnd_1}$ に含まれるが $\in \overline{\mathcal{L}}^{cnd_2}$ には含まれないオブジェクト o_1 と $\overline{\mathcal{L}}^{cnd_2}$ に含まれるオブジェクト o_2 に対し，cnd_2 は評価基準なので $o_1 < o_2$ である．よって，o_1 は cnd_1 を満たすが cnd_2 は満たさず，o_2 は cnd_2 を満たすので $cnd_1 < cnd_2$ である．

$\overline{\mathcal{L}}^{cnd_2} \subseteq \overline{\mathcal{L}}^{cnd_1}$ でなければ，$o_2 \in \overline{\mathcal{L}}^{cnd_2}$ かつ $o_2 \notin \overline{\mathcal{L}}^{cnd_1}$ であるオブジェクト o_2 が存在する．o_2 は cnd_2 を満たし cnd_1 を満たさない．一方，$o_1 \in \overline{\mathcal{L}}^{cnd_1}$ かつ $o_1 \notin \overline{\mathcal{L}}^{cnd_2}$ であるオブジェクト o_1 に対し，cnd_1 は評価基準なので $o_2 < o_1$ である．o_2 は cnd_2 を満たし o_1 は cnd_2 を満たさず cnd_1 を満たすが $o_1 < o_2$ ではないので，$cnd_1 < cnd_2$ ではない．したがって，$cnd_1 < cnd_2$ ならば $\overline{\mathcal{L}}^{cnd_2} \subseteq \overline{\mathcal{L}}^{cnd_1}$ である． （証明終）

ラベル集合 $\mathcal{L}$ と評価基準 cnd_1，cnd_2 に対し，評価基準 cnd_1 が評価基準 cnd_2 を含意する（$cnd_1 \Rightarrow cnd_2$）とき，$\overline{\mathcal{L}}^{cnd_2} \subseteq \overline{\mathcal{L}}^{cnd_1}$ である．$\overline{\mathcal{L}}^{cnd_2} \subseteq \overline{\mathcal{L}}^{cnd_1}$ ならば $cnd_1 \Rightarrow cnd_2$ なので，包含関係と含意関係は等価である．補題 1 より評価基準の強さの関係は評価基準を満たすオブジェクト集合の包含関係で決まるので，cnd_1 と cnd_2 に強さの関係があるかは cnd_1 と cnd_2 に含意関係があるかどうかで判断できる．

［定理 1］ 評価基準 cnd_1，cnd_2 に対し，$cnd_1 < cnd_2$ と $cnd_2 \Rightarrow cnd_1$ は等価である．

（証明） $cnd_1 < cnd_2$ ならば，補題 1 より $\overline{\mathcal{L}}^{cnd_2} \subseteq \overline{\mathcal{L}}^{cnd_1}$ である．$\overline{\mathcal{L}}^{cnd_2} \subseteq \overline{\mathcal{L}}^{cnd_1}$ ならば，cnd_2 を満たすオブジェクトは cnd_1 も満たすので $cnd_2 \Rightarrow cnd_1$ となる．また，$cnd_2 \Rightarrow cnd_1$ ならば $\overline{\mathcal{L}}^{cnd_2} \subseteq \overline{\mathcal{L}}^{cnd_1}$ であり，補題 1 より $\overline{\mathcal{L}}^{cnd_2} \subseteq \overline{\mathcal{L}}^{cnd_1}$ ならば $cnd_1 < cnd_2$ である． （証明終）

定理 1 より，評価基準の強さの関係はその含意で考えることができる．評価基準 LE，LA，LN はランク付集合ラベル全体を対象にしているのに対し，PE，PA，PN は主ラベルの集合を対象にしている．主ラベルの集合はランク付集合ラベルの部分集合なので，PE，PA，PN はそれぞれ LE，LA，LN を含意し，強さの関係がある．

［補題 2］ 評価基準 LE と PE，LA と PA，PN と LN に対し，$LE < PE$，$LA < PA$，$PN < LN$ である．

（証明） PE を満たすオブジェクト o では $Ring_{\mathcal{L}}(P(o)) \neq \phi$ である．$P(o) \subseteq L(o)$ なので $Ring_{\mathcal{L}}(L(o)) \neq \phi$ であり，o は LE を満たす．よって，$PE \Rightarrow LE$ なので $LE < PE$ である．同様に，$Red_{P(o)}(\mathcal{L}) = \mathcal{L}$ ならば $Red_{L(o)}(\mathcal{L}) = \mathcal{L}$，$Ring_{\mathcal{L}}(L(o)) = L(o)$ ならば $Ring_{\mathcal{L}}(P(o)) = P(o)$ であり，それぞれ $PA \Rightarrow LA$，$LN \Rightarrow PN$ が成り立つので，$LA < PA$，$PN < LN$ である． （証明終）

評価基準 PN を満たすオブジェクト o の主ラベルは $\mathcal{L}$ と無関係なラベルに関連しないが，o は主ラベルを必ず持つので $\mathcal{L}$ に関連する．よって，o は PE を満たすので，2 つの評価基準には含意関係があり，強さの関係がある．

［補題 3］ 評価基準 PE と PN に対し，$PE < PN$ である．

（証明） PN を満たすオブジェクト o では $Ring_{\mathcal{L}}(P(o)) = P(o)$ であり，$\mathcal{L}$ と無関係なラベルに関連するラベルを $P(o)$ に含まない．$P(o) \neq \phi$ なので，$P(o)$ 中に $\mathcal{L}$ に関連するラベルが必ず存在するため $Ring_{\mathcal{L}}(P(o)) \neq \phi$ であり，o は PE を満たす． （証明終）

評価基準 LA を満たすオブジェクト o は，ラベル集合 $\mathcal{L}$ のすべてのラベルに関連するので LE も満たす．PE と PA についても同様であり，強さの関係がある．

［補題 4］ 評価基準 LE と LA，PE と PA に対し，$LE < LA$，$PE < PA$ である．

（証明） LA を満たすオブジェクト o は，$Red_{L(o)}(\mathcal{L}) = \mathcal{L}$ であり，$\mathcal{L} \neq \phi$ より $Red_{L(o)}(\mathcal{L}) \neq \phi$ なので LE を満たす．よって，$LA \Rightarrow LE$ なので，$LE < LA$ である．$PE < PA$ についても同様である． （証明終）

補題 2，3，4 で示した評価基準の強さの関係は，評価基準 cnd_i と cnd_j が $cnd_i < cnd_j$ であることを cnd_j から cnd_i への矢印で表すと，図 2 となる．例えば，評価基準 LN から PN への矢印は $PN < LN$ を示している．

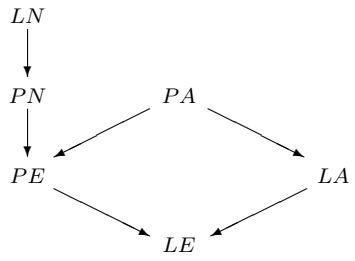


図 2 評価基準の強さの関係

評価基準の強さの関係には，補題 2，3，4 で得られた評価基準の強さの関係から推移律で得られるもの以外は存在しないことは容易に示せる．例えば，オブジェクト o は LA を満たしても，$\mathcal{L}$ に関連するすべてのラベルが副ラベルの場合は $Ring_{\mathcal{L}}(P(o)) \neq \phi$ ではないため，PE を満たすとは限らない．

4 評価基準の導出

2 節の評価基準から新たな評価基準を導出できる場合がある．例えば，評価基準 LA と PE の両方の条件とする $LA \wedge PE$ は 2 節の評価基準とは異なる．本節では，これまでの評価基準から導出される評価基準について述べる．

評価基準 cnd_1 と cnd_2 の積を $cnd_1 \wedge cnd_2$ とする．オブジェクト o_1 が $cnd_1 \wedge cnd_2$ を満たさなければ，o_1 は少なくとも cnd_1 と cnd_2 のどちらかは満たさない．o_1 が cnd_1 を満たさないとき，$cnd_1 \wedge cnd_2$ を満たすオブジェクト o_2 は cnd_1 を満たすので $o_1 < o_2$ である．よって，o_2 が $cnd_1 \wedge cnd_2$ を満たし o_1

が $cnd_1 \wedge cnd_2$ を満たさないとき，$o_1 < o_2$ なので $cnd_1 \wedge cnd_2$ は評価基準である．

評価基準の集合を $\boldsymbol{C} = \{cnd_1, cnd_2, \cdots, cnd_n\}$ で，$\boldsymbol{C}$ から導出される評価基準集合，すなわち，$\boldsymbol{C}$ の要素の積によって得られる評価基準集合を $\boldsymbol{C}^+ = \{cnd_1, cnd_2, \cdots, cnd_n, cnd_1 \wedge cnd_2, \cdots, cnd_1 \wedge cnd_2 \wedge \cdots \wedge cnd_n\}$ とする．評価基準集合 $\boldsymbol{C}_1$ と $\boldsymbol{C}_2$ は，$\boldsymbol{C}_1^+ = \boldsymbol{C}_2^+$ のとき等価であるという．また，評価基準集合 $\boldsymbol{C}$ は，等価な真部分集合が存在しないとき基底であるという．

［補題 5］ 評価基準 LE，LA，LN，PE，PA，PN を要素とする評価基準集合は基底である．

（証明） 評価基準集合 $\boldsymbol{C} = \{LE, LA, LN, PE, PA, PN\}$ とその任意の評価基準 cnd（$\in \boldsymbol{C}$）に対し，$cnd \notin (\boldsymbol{C} - cnd)^+$ なので，$\boldsymbol{C}$ の真部分集合は $\boldsymbol{C}$ と等価ではない．よって，$\boldsymbol{C}$ は基底である． （証明終）

補題 5 より評価基準集合 $\boldsymbol{C}_B = \{LE, LA, LN, PE, PA, PN\}$ は基底であり，$\boldsymbol{C}_B$ から導出される評価基準集合 $\boldsymbol{C}_B^+$ について検討する．

評価基準 cnd_i と cnd_j（$i \neq j$）は，$cnd_i \Rightarrow cnd_j$ かつ $cnd_j \Rightarrow cnd_i$ であるとき同値であるといい，$cnd_i \Leftrightarrow cnd_j$ で表す．$cnd_i \Leftrightarrow cnd_j$ であれば，cnd_i を満たすオブジェクト集合と cnd_j を満たすオブジェクト集合は一致する（$\overline{\mathcal{L}}^{cnd_i} = \overline{\mathcal{L}}^{cnd_j}$）．

評価基準 cnd_i と cnd_j に対し，$cnd_i \Rightarrow cnd_j$ ならば $cnd_i \Leftrightarrow cnd_i \wedge cnd_j$ なので，$cnd_i \wedge cnd_j$ は cnd_i と同値であり新たな評価基準ではない．図 3 は，基底 $\boldsymbol{C}_B$ の評価基準の積で得られる評価基準を含意の関係で整理したものである．例えば，$PE \Rightarrow LE$ なので，$PE \Leftrightarrow LE \wedge PE$ であり，$LE \wedge PE$ は新たな評価基準ではない．下線の評価基準は同値である評価基準が $\boldsymbol{C}_B$ に存在せず，新たな評価基準である．

	LE	PE	LA	PA	LN	PN
LE	–	PE	LA	PA	LN	PN
PE	–	–	$\underline{LA \wedge PE}$	PA	LN	PN
LA	–	–	–	PA	$\underline{LA \wedge LN}$	$\underline{LA \wedge PN}$
PA	–	–	–	–	$\underline{PA \wedge LN}$	$\underline{PA \wedge PN}$
LN	–	–	–	–	–	LN
PN	–	–	–	–	–	–

図 3 基底の評価基準の積で得られる評価基準

図 3 の新たに得られた評価基準と基底 $\boldsymbol{C}_B$ の評価基準との積で得られる評価基準には新たな評価基準は存在しない（図 4）．

したがって，オブジェクトのラベル集合への関連の強さを評価するための基準は，基底 $\boldsymbol{C}_B = \{LE, LA, LN, PE, PA, PN\}$ 及び $LA \wedge PE$，$LA \wedge LN$，$LA \wedge PN$，$PA \wedge LN$，$PA \wedge PN$ である．

5 評価基準系列

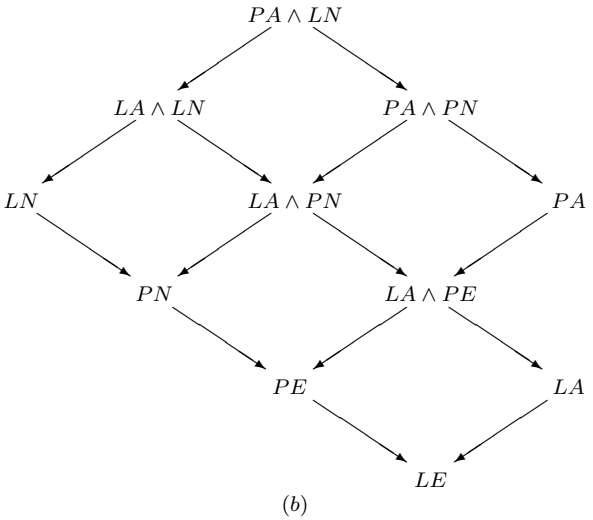
ラベル集合 $\mathcal{L}$ が与えられとき，オブジェクト o が満たす評価基準によって o と $\mathcal{L}$ との関連の強さを判断できるので，o が満

	$LA \wedge PE$	$LA \wedge LN$	$LA \wedge PN$	$PA \wedge LN$	$PA \wedge PN$
LE	$LA \wedge PE$	$LA \wedge LN$	$LA \wedge PN$	$PA \wedge LN$	$PA \wedge PN$
PE	–	$LA \wedge LN$	$LA \wedge PN$	$PA \wedge LN$	$PA \wedge PN$
LA	–	–	–	$PA \wedge LN$	$PA \wedge PN$
PA	$PA \wedge PE$	$PA \wedge LN$	$PA \wedge PN$	–	–
LN	$LA \wedge LN$	–	$LA \wedge LN$	–	$PA \wedge LN$
PN	$LA \wedge PN$	$LA \wedge LN$	–	$PA \wedge LN$	–

図 4 基底から得られた評価基準と基底との積で得られる評価基準

たす最も強い評価基準が o と $\mathcal{L}$ との関連の強さとみなすことができる．本節は，これまでに導出した評価基準の強さの順序を明らかにすることで，強さの比較が可能な評価基準の集合を系列として定義し，関連の強さを段階的かつ一元的に評価できる理論的枠組みを与える．

評価基準 cnd_i，cnd_j，cnd_k に対し，$cnd_i \wedge cnd_j \Rightarrow cnd_i$ なので $cnd_i < cnd_i \wedge cnd_j$ であり，$cnd_j \Rightarrow cnd_k$ ならば $cnd_i \wedge cnd_j \Rightarrow cnd_i \wedge cnd_k$ となり $cnd_i \wedge cnd_k < cnd_i \wedge cnd_j$ である．例えば，$LA \wedge PE \Rightarrow LA$ なので $LA < LA \wedge PE$ であり，$PN \Rightarrow PE$ なので $LA \wedge PN \Rightarrow LA \wedge PE$ となり $LA \wedge PE < LA \wedge PN$ である．また，評価基準 PA を満たすオブジェクト o は $Red_{P(o)}(\mathcal{L}) = \mathcal{L}$ なので，$Red_{L(o)}(\mathcal{L}) = \mathcal{L}$ であり LA を満たし，$Red_{P(o)}(\mathcal{L}) \neq \phi$ であり PE を満たす．よって，$PA \Rightarrow LA \wedge PE$ であり $LA \wedge PE < PA$ である．これらから，評価基準の強さの関係は図 5 となる．



(b)

図 5 導出された評価基準の強さの関係

図 5 の評価基準 $PA \wedge LN$ から LE までの経路は，評価基準の強さの順序を示している．最も弱い LE と最も強い $PA \wedge LN$ を省略すると次のものとなる．

I: $LA < LA \wedge PE < PA < PA \wedge PN$

II: $LA < LA \wedge PE < LA \wedge PN < PA \wedge PN$

III: $LA < LA \wedge PE < LA \wedge PN < LA \wedge LN$

IV: $PE < LA \wedge PE < PA < PA \wedge PN$

V: $PE < LA \wedge PE < LA \wedge PN < PA \wedge PN$

VI: $PE < LA \wedge PE < LA \wedge PN < LA \wedge LN$

VII: $PE < PN < LA \wedge PN < PA \wedge PN$

$VIII$: $PE < PN < LA \wedge PN < LA \wedge LN$

IX: $PE < PN < LN < LA \wedge LN$

評価基準 cnd_i，cnd_j に強さの関係があれば，cnd_i と cnd_j は比較可能であり，そのような評価基準集合は共通の体系の下での評価基準である．

［定義 3］ 評価基準集合 $\boldsymbol{C}$ に対し，$\forall cnd_i, cnd_j \in \boldsymbol{C}, cnd_i < cnd_j$ または $cnd_j < cnd_i$ であるとき，$\boldsymbol{C}$ は評価基準系列あるいは系列という．

図 5 の $PA \wedge LN$ から LE までの経路に含まれる評価基準は比較可能なので，経路の評価基準集合及びその部分集合は系列である．例えば，経路 I の $\boldsymbol{C}_I = \{LE, LA, LA \wedge PE, PA, PA \wedge PN, PA \wedge LN\}$ 及びその部分集合は系列である．

［定理 2］ 評価基準集合 $\boldsymbol{C}$ が系列であることは，$\boldsymbol{C}$ が評価基準 $PA \wedge LN$ から LE までの経路の評価基準集合の部分集合であることと等価である．

（証明） 評価基準集合 $\boldsymbol{C}$ が評価基準 $PA \wedge LN$ から LE までの経路の評価基準集合の部分集合であれば，$\boldsymbol{C}$ に含まれる評価基準は比較可能なので $\boldsymbol{C}$ は系列である．評価基準集合 $\boldsymbol{C}$ における任意の 2 つの評価基準が $PA \wedge LN$ から LE までの同じ経路に含まれなければ，それらの評価基準は比較可能ではないので $\boldsymbol{C}$ は系列ではない． （証明終）

分析対象をラベル集合 $\mathcal{L}$ で与えたときに，どのようなオブジェクトを分析範囲に含めるのかという分析対象の解釈には様々な考え方がある．$\mathcal{L}$ の解釈は複数存在し，評価基準はそれらに対応する．$\mathcal{L}$ の解釈には，

α: $L(o)$ に $\mathcal{L}$ に関するラベルがある，すなわち LE を満たす，

β: $P(o)$ に $\mathcal{L}$ に関するラベルがある，すなわち PE を満たす，

γ: $L(o)$ に $\mathcal{L}$ に関するすべてのラベルがある，すなわち LA を満たす，

δ: $P(o)$ に $\mathcal{L}$ に関するすべてのラベルがある，すなわち PA を満たす，

ϵ: $L(o)$ に $\mathcal{L}$ 以外に関するラベルがない，すなわち LN を満たす，

ζ: $P(o)$ に $\mathcal{L}$ 以外に関するラベルがない，すなわち PN を満たす，

η: $L(o)$ に $\mathcal{L}$ に関して過不足なくラベルがある，すなわち $LA \wedge LN$ を満たす，

θ: $P(o)$ に $\mathcal{L}$ に関して過不足なくラベルがある，すなわち $PA \wedge PN$ を満たす

ときオブジェクト o は $\mathcal{L}$ の分析範囲に含まれるという解釈が考えうる．$\mathcal{L}$ の解釈は，$L(o)$ と $P(o)$ のどちらを対象にするのかという視点と，$\mathcal{L}$ に関するラベル，$\mathcal{L}$ 以外に関するラベル，その両方，すなわち $\mathcal{L}$ に関して過不足ないラベルのどれを対象にするのかという視点がある．$\mathcal{L}$ の解釈をこれらの視点で整理したものが図 6 である．

	$L(o)$	$P(o)$
$\mathcal{L}$ に関するラベル	$\alpha : LE$ $\gamma : LA$	$\beta : PE$ $\delta : PA$
$\mathcal{L}$ 以外に関するラベル	$\epsilon : LN$	$\zeta : PN$
$\mathcal{L}$ に関して過不足ないラベル	$\eta : LA \wedge LN$	$\theta : PA \wedge PN$

図 6 ラベル集合 $\mathcal{L}$ の解釈

系列 I から IX の特徴は，$\mathcal{L}$ の解釈に基づいて説明できる（図 7）．分析目的に応じて $\mathcal{L}$ の解釈が決まり，対応する評価基準を満たすオブジェクト集合が分析範囲になる．例えば，$\mathcal{L} =$ { 日本, 中国 } のつながりを分析する目的であれば，$\mathcal{L}$ は γ や δ のように $L(o)$ や $P(o)$ に $\mathcal{L}$ に関するすべてのラベルがあるオブジェクトが分析範囲に含まれると解釈され，評価基準 LA や PA を満たすオブジェクト集合が分析範囲になる．

$\mathcal{L}$ に関するラベルがないオブジェクトは一般に $\mathcal{L}$ の分析範囲には含まれないので，$\mathcal{L}$ の分析範囲に含まれるオブジェクトは $\mathcal{L}$ に関するラベルがある．よって，$\mathcal{L}$ を α で解釈する，すなわち LE を満たすオブジェクト集合が $\mathcal{L}$ の分析範囲として最大範囲になる．一方，$\mathcal{L}$ の最小範囲は，最も強い評価基準 $PA \wedge LN$ を満たすオブジェクト集合である．系列 I から IX は，LE から $PA \wedge LN$ までの経路に含まれる評価基準集合であり，共通する LE と $PA \wedge LN$ を除いた評価基準集合が異なり，系列の特徴はそれらの評価基準集合によるものである．

図 6 から，$\mathcal{L}$ に関するラベルを対象にする解釈には，α を除くと β，γ，δ があるが，δ は β と γ を含意する．同様に，$\mathcal{L}$ 以外に関するラベルを対象にする解釈には，ϵ と ζ があるが，ϵ は ζ を含意する．他の解釈を含意しない β，γ，ζ は起点となる解釈といえ，系列 I から IX は LA，PE，PN のいずれかの評価基準が起点になる．$\mathcal{L}$ に関して過不足ないラベルを対象にする解釈は，η と θ であり，互いに含意せず，それぞれ $LA \wedge LN$ と $PA \wedge PN$ が対応する．η，θ は β，γ，ζ を含意するので，η と θ は終点となる解釈といえる．LA，PE，PN が起点となり，$LA \wedge LN$ もしくは $PA \wedge PN$ を目指す．その際に，PA，PN，LA，LN のどれを優先させるかで分類できる．

起点となる 解釈:評価基準	優先する 解釈:評価基準	終点となる 解釈:評価基準	対応する 系列
$\gamma : LA$	$\delta : PA$	$\theta : PA \wedge PN$	I
$\gamma : LA$	$\zeta : PN$	$\theta : PA \wedge PN$	II
$\gamma : LA$	$\epsilon : LN$	$\eta : LA \wedge LN$	III
$\beta : PE$	$\delta : PA$	$\theta : PA \wedge PN$	IV
$\beta : PE$	$\gamma : LA$	$\theta : PA \wedge PN$	V
$\beta : PE$	$\zeta : PN$	$\theta : PA \wedge PN$	(VII)
$\beta : PE$	$\gamma : LA$	$\eta : LA \wedge LN$	VI
$\beta : PE$	$\epsilon : LN$	$\eta : LA \wedge LN$	(IX)
$\zeta : PN$	$\delta : PA$	$\theta : PA \wedge PN$	VII
$\zeta : PN$	$\gamma : LA$	$\eta : LA \wedge LN$	$VIII$
$\zeta : PN$	$\epsilon : LN$	$\eta : LA \wedge LN$	IX

図 7 系列の特徴

6 系列を用いたデータ分析

評価基準系列は関連の強さを比較できる評価基準集合なので，比較可能な基準を提示し詳細な分析を支援する．系列に含まれる評価基準を段階的に用いることで，分析対象に対する関連の強さの差異を利用した分析ができる．一方，同一系列にない評価基準を用いてデータ集合を比較しても，比較した結果は評価基準の条件の差で説明できないので，関連の強さの差異を利用した分析ができない．本節は，分析シナリオの提示や人工データを用いたシミュレーションにより，データ分析における系列の利用法と有用性について述べる．

企業データを対象とする分析シナリオ　企業における事業の多角化が業績に与える影響で，自動車産業と情報技術産業のつながりを分析する．データを集約し評価する一般の分析では，自動車と情報技術の両方に関連する企業の利益の平均値を全体の平均値と比較することが行われる．これは，分析対象をラベル集合 $\mathcal{L}=\{$ 自動車，情報技術 $\}$ で与えて，$\mathcal{L}$ を γ で解釈，すなわち $\mathcal{L}$ に関してすべてのラベルがあるオブジェクト集合（$\overline{\mathcal{L}}^{LA}$）の集約値を全オブジェクト集合の集約値と比較することに対応する．ランク付集合ラベルが与えられていれば，$\mathcal{L}$ を δ で解釈し評価基準 PA で集約を行うことで $\mathcal{L}$ との関連が強いオブジェクト集合と比較でき，詳細な情報を得ることができる．例えば，$\overline{\mathcal{L}}^{LA}$ よりも $\overline{\mathcal{L}}^{PA}$ の平均値の方が高ければ，両業種に強く関連する企業の方が利益が高いといえ，両業種のつながりが強く親和性や相乗効果が高いと推測できる．このような分析は，起点となる解釈を γ，優先する解釈を δ とする系列 I に基づく分析である．また，系列 I における終点となる解釈は θ であり評価基準 $PA\wedge PN$ が対応するので，他業種の影響について分析が可能である．$\overline{\mathcal{L}}^{PA\wedge PN}$ の平均値が高ければ，両業種へ特化することの効果が推測できる．

系列は様々な分析目的に対応する．例えば，自動車と情報技術に関連する企業に対し，両業種への特化を優先して分析する場合には，$\mathcal{L}$ を ϵ と解釈し，対応する LN を優先する系列 III を用いればよい．系列 III には終点の $LA\wedge LN$ よりも 1 段階弱い評価基準 $LN\wedge PN$ が含まれるので，$\overline{\mathcal{L}}^{LA}$ と $\overline{\mathcal{L}}^{LA\wedge PN}$ との比較，$\overline{\mathcal{L}}^{LA\wedge PN}$ と $\overline{\mathcal{L}}^{LA\wedge LN}$ との比較をすることで詳細な分析が可能になる．$\overline{\mathcal{L}}^{LA\wedge PN}$ の利益の平均値の方が $\overline{\mathcal{L}}^{LA}$ よりも高ければ，両業種に事業展開している企業のうち，他業種を主としていない方が利益が高い．さらに，$LA\wedge LN$ により他業種に事業展開していない企業との比較ができるので，両業種へ特化する効果について詳細な分析ができる．

一方，同一系列にない評価基準，例えば $\overline{\mathcal{L}}^{PA}$ と $\overline{\mathcal{L}}^{LA\wedge LN}$ の集約値を比較し $\overline{\mathcal{L}}^{PA}$ の方が高かったとしても，両業種に強く関連していることと両業種に特化していることのどちらの要因なのか分からない．このように同一系列にない PA と $LA\wedge LN$ による比較では，$\overline{\mathcal{L}}^{PA}$ の集約値の方が高いということは分かるがそれ以上の知見は得られない．

ラベル集合 $\mathcal{L}$ の要素が単数のときは $\mathcal{L}$ に対する特化の程度に関する分析のみができる．$\mathcal{L}$ のすべてのラベルに関連するという評価基準の LA は LE と同値であり，$\mathcal{L}$ のすべてのラベルに強く関連するという評価基準の PA は PE と同値である．図 5 において LA と PA 及びそれらから導出される評価基準は LE，PE，PN，LN のいずれかと同値である．よって，ラベル集合の要素が単数のとき，系列は評価基準集合 $\boldsymbol{C}=\{LE,PE,PN,LN\}$ 及びその部分集合になる．

教育と給与データを対象とする分析シナリオ　教育と給与に関するデータを対象に高等教育における専攻分野と給与水準との関係を分析する．個人の専攻分野を 1 件のオブジェクトとし，学位や就学期間といった情報に基づきランク付集合ラベルが与えられているものとする．ラベル集合 $\mathcal{L}=\{$ 情報工学 $\}$ と系列 $\boldsymbol{C}=\{LE,PE,PN,LN\}$ を用いて，評価基準 LE から LN を満たす集団になる程その集団の給与水準が向上していれば，情報工学に特化して専攻したことが給与水準の向上につながる可能性があることが示唆される．ラベルを “経済学” に変えれば，経済学の専攻と給与水準に関する知見が得られる．

人工的な観光データによるシミュレーション　オブジェクトは顧客が観光した施設を表すランク付集合ラベルと顧客満足度を示す数値からなるものとし，100 万件のオブジェクトを人工的に生成し系列を用いて分析する．A，B，C の 3 地区を想定し，ラベル集合 $\mathcal{L}=\{$A 地区，B 地区 $\}$ における行動と満足度の関係に注目する．A 地区には $a1$，$a2$，B 地区には $b1$，$b2$，C 地区には $c1$，$c2$ という観光施設があるものとする．a_1 から c_2 の集合を $K=\{a1,a2,b1,b2,c1,c2\}$ とすると，オブジェクト o_i $(1\leqq i\leqq 1,000,000)$ のランク付集合ラベル $L(o_i)$ は，$L(o_i)\subseteq K$，$P(o_i)\neq\phi$ である．

施設には 5 段階の評判（5S:良い，4S:やや良い，3S:普通，2S:やや悪い，1S:悪い）があるものとし，顧客が施設を観光する確率，すなわち，主ラベル，副ラベルが付される確率はその施設の評判に応じて決める．観光施設 k（$\in K$）の評判が良いほどで高い確率で主ラベルが付されるものとし，“5S:良い” 評判では 45%とし，評判が 1 段階下がるごとに確率は 5%下がるものとする．副ラベルについては評判によらず 30%とする．すなわち，k が “5S:良い” 評判であれば，主ラベルが 45%，副ラベルが 30%，ラベルが付されない確率が 25%である．各施設の評判は前提条件として任意に設定し，評判に応じた確率分布に基づいて施設ごとにラベルを選ぶことで $L(o_i)$ を生成する．

o_i の顧客満足度 $V(o_i)$ は，$L(o_i)$ と各施設の評判から生成する．$V(o_i)$ は o_i の顧客が観光した施設 k の評価値 $E(o_i,k)$（$E(o_i,k)\in\{2,1,0,-1,-2\}$，2:良い，1:やや良い，0:普通，-1:やや悪い，-2:悪い）を合算した値とする．施設の評判が良いほど高い確率で良い評価値が与えられる確率分布を設定し，その分布に基づいて選んだ値が $E(o_i,k)$ になる．観光した施設の評判が “5S:良い” 場合では，評価値が 2 になるのは 60%，1 は 30%，0 は 5%，-1 は 3%，-2 は 2%という確率分布を設定している．また，訪問した施設のラベルのランクによって満足度に与える影響が異なると考え，副ラベルを 1 とする主ラベルの重み W を 2 とする．$V(o_i)=\sum_{k\in P(o_i),k'\in S(o_i)}W*E(o_i,k)+E(o_i,k')$ であり，$P(o_i)=K$ かつ $\forall k\in K$，$E(o_i,k)=2$（$E(o_i,k)=-2$）のとき，$V(o_i)=24$（$V(o_i)=-24$）で最大値 (最小値) になる．

A 地区，B 地区にある施設を主に観光した顧客を対象に，両方の地区を主に観光した顧客の満足度を分析する．分析目的に対応する系列は，$\mathcal{L}$ を β で解釈することが起点で δ の解釈を優先する系列 IV である．観光施設の評判を 3 通り設定し，人工データを生成して分析した結果は表 1 の通りである．

ケース 1 において，PE を満たす集団の満足度の平均値が 3.87 であるのに対し，$LA \wedge PE$ では 4.08，PA では 4.63 と段階的に高くなる．評価基準の条件の差で説明すると，A 地区と B 地区の両方の施設を主に観光した顧客ほど満足度が高いことを意味している．これは A 地区と B 地区の両方に評判の良い施設があるという前提条件によるものである．ケース 2 において，PE の 1.30 から $LA \wedge PE$ では 0.88，PA では 0.49 と段階的に低くなるので，A 地区と B 地区の両方の施設を主に観光した顧客ほど満足度が低いといえる．これは A 地区の施設の評判は良いが，B 地区の施設の評判は悪いためである．

系列 IV には $PA \wedge PN$ や $PA \wedge LN$ といった評価基準が含まれるので，C 地区の施設を観光した影響を考慮した分析もできる．ケース 3 において，PA の 2.34 から $PA \wedge PN$ では 3.48，$PA \wedge LN$ では 4.62 と高くなるので，C 地区にある施設を主として観光してない，さらに全く観光していない顧客の方が満足度が高いといえる．これは A 地区や B 地区の施設よりも C 地区の施設が評判が悪いためである．

$\mathcal{L}$ への関連の強さを段階的に高めて分析範囲を特定し集約値を検証した．集約値の変化は評価基準の条件の差で説明でき，$\mathcal{L}$ に関する知見が得られる．その知見は観光施設の評判という前提条件に合致し，系列を用いた分析の正しさを示している．

表 1　系列 IV を用いた人工データの分析結果

	ケース 1		ケース 2		ケース 3	
評価基準		満足度		満足度		満足度
	個数	平均値	個数	平均値	個数	平均値
全体	1,000,000	3.70	1,000,000	1.20	1,000,000	1.36
LE	995,411	3.71	992,071	1.21	996,482	1.38
PE	992,044	3.87	894,297	1.30	939,993	1.58
$LA \wedge PE$	802,704	4.08	712,364	0.88	817,936	1.79
PA	435,477	4.63	328,926	0.49	444,015	2.34
$PA \wedge PN$	183,472	4.64	139,168	0.48	249,297	3.48
$PA \wedge LN$	53,369	4.64	40,290	0.47	89,894	4.62
標準偏差		3.47		3.69		3.85
最大値		21		20		18
最小値		-16		-17		-17
観光施設	(a1,a2,b1,b2,c1,c2)		(a1,a2,b1,b2,c1,c2)		(a1,a2,b1,b2,c1,c2)	
評判	(5S,4S,5S,2S,3S,3S)		(5S,5S,1S,1S,3S,3S)		(5S,4S,5S,2S,1S,1S)	

7 おわりに

本論文は，ラベル集合とランク付集合ラベルデータとの関連の強さを検討し，その関連の強さを評価するための基準は，基底 $\boldsymbol{C}_B = \{LE, LA, LN, PE, PA, PN\}$ 及び基底の組合せから得られる $LA \wedge PE$，$LA \wedge LN$，$LA \wedge PN$，$PA \wedge LN$，$PA \wedge PN$ であることを示した．評価基準を満たすデータを分析範囲にすることで，分析対象と分析範囲の関係が明確になるので，関連の強さの差異に基づいて分析結果を適切に検証できる．さらに，評価基準間の含意から強さの順序を明らかにすることで系列を導いた．

分析対象はラベル集合で与えられ，データが系列の評価基準をどこまで満たすかによってラベル集合との関連の強さが段階的かつ一元的に定まる．関連の強さが異なるデータ集合を比較することで，分析対象に関する情報を得られる．

系列には様々なものがあるので，多様な分析目的に対応できる．また，系列によって比較可能な評価基準を提示できるので，データの分析範囲を多面的な視点で特定して分析することを支援する．

本論文では主ラベルと副ラベルの区分で議論したが，区分の数を増やすことでより詳細な分析が可能になる．本手法は区分の数を増やしたランクにも拡張可能である．

文　献

[1] Bi, W. and Kwok, J.: Multi-Label Classification on Tree- and DAG-Structured Hierarchies, *Proc. Int'l Conf. on Machine Learning* (ICML '11), pp. 17–24 (2011).

[2] 古川哲也，葛西正裕: 集合ラベルを持つデータの集約範囲の記述, 情報処理学会論文誌：データベース, 情報処理学会, Vol. 3, No. 3, pp. 11-19 (2010).

[3] Hu, H., Wen, Y., Chua, T., and Li, X.: Toward Scalable Systems for Big Data Analytics: A Technology Tutorial, *IEEE Access*, Vol. 2, pp. 652–687 (2014).

[4] Kuzunishi, M. and Furukawa, T.: Strength of Relationship Between Multi-labeled Data and Labels, *Proc. Information and Communication Technology - Third IFIP TC 5/8 Int'l Conf., ICT-EurAsia 2015, and 9th IFIP WG 8.9 Working Conf., CONFENIS 2015, Held as Part of WCC 2015*, pp. 99–108 (2015).

[5] Ren, Z., Peetz, M., Liang, S., Dolen, W., and Rijke, M.: Hierarchical Multi-Label Classification of Social Text Streams, *Proc. ACM Int'l Conf. on Research and Development in Information Retrieval* (SIGIR '14), pp. 213–222 (2014).

[6] Rouse, J., Sauuders, C., Szedmark, S., and Shawe-Taylor, J.: Kernel-Based Learning of Hierarchical Multilabel Classification Models, *Journal of Machine Learning Research 7*, pp. 1601–1626 (2006).

[7] Sinaeepourfard, A., Garcia, J., Masip-Bruin, X., and Marin-Tordera, E.: Towards a Comprehensive Data LifeCycle Model for Big Data Environments, *Proc. ACM 3rd Int'l Conf. Big Data Computing Applications and Technologes* (BDCAT '16), pp. 100–106 (2016).

[8] Silla, C. and Freitas, A.: A survey of hierarchical classification across different application domains, *Data Mining and Knowledge Discovery*, Vol. 22, Issue. 1-2, pp. 311–72 (2011).

[9] Steinhauer, J., Delcambre, L., Maier, D., Lykke, M., and Tran,V.: Tags in Domain-Specific Sites - New Information?, *Proc. the 11th Annual Int'l ACM/IEEE joint Conf. on Digital Libraries* (JCDL '11), pp. 109–112 (2011).

[10] Tang, B., Han, S., Yiu, M., Ding, R., and Zhan, D.: Extracting Top-K Insights from Multi-dimensional Data, *Proc. IEEE/ACM Conf. on Management of Data* (SIGMOD '17), pp. 1509–1524 (2017).

[11] Wasay, A., Wei, X., Dayan, N., and Idreos, S.: Data Canopy: Accelerating Exploratory Statistical Analysis, *Proc. ACM SIGMOD Int'l Conf. on Management of Data* (SIGMOD '17), pp. 557–572 (2017).

[12] Wu, F., Wang, Z., Zhang, Z., Yang, Y., Luo, J., Zhu, W., and Zhuang, Y.: Weakly Semi-Supervised Deep Learning for Multi-Label Image Annotation, *IEEE TRANSACTIONS ON BIG DATA*, Vol. 1, No. 3, pp. 109–122 (2015).

[13] Zhu, X., Song, S., Lian, X., Wang, J., and Zou, L.: Matching Heterogeneous Event Data, *Proc. ACM SIGMOD Int'l Conf. on Management of Data* (SIGMOD '14), pp. 1211–1222 (2014).