3B-04

ランク付集合ラベルデータのための評価基準系列

葛西正裕 †
愛知学院大学経済学部 †

古川哲也 ‡
九州大学大学院経済学研究院 ‡

1 はじめに

データを検索や分析で利用するために，データには概念階層（シソーラス）などに基づいて該当する複数のカテゴリのラベル（集合ラベル）が付される [1]．一般には，各カテゴリとデータの関連の強さは異なるので，関連の強さを区分するランクを導入する．ラベル集合を用いて関連の強いデータを検索したりラベル集合との関連の強さに基づく分析を行うためには，データのラベル集合への関連の強さを明確にしておく必要がある．本稿では，ラベル集合とランク付集合ラベルとの関連の強さを評価する基準を導出する．さらに，評価基準間の強さの順（系列）を明らかにすることで，関連の強さを段階的かつ一元的に評価するための理論的な枠組みを示す．

2 ランクの導入と集合ラベルの性質

データに付された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，例えば，業種という属性では製造業，輸送機器，四輪自動車といった上下関係がある．本稿では，データには単一属性のもとで最下層のカテゴリ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 o，ラベルを L，ラベル集合を $\boldsymbol{L} = \{L_1, L_2, \cdots, L_n\}$ とし，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．ラベルの概念の上下関係を $\prec$ とする．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 の下位であり，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．L が $\boldsymbol{L}$ 中のいずれかのラベルに関連するとき，L は $\boldsymbol{L}$ に関連するという．

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

ラベル集合 $\mathcal{L}$（$\mathcal{L} \neq \phi$）とオブジェクト o に対し，$\mathcal{L}$ に関連するラベルが $L(o)$ にある，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．

オブジェクト o の集合ラベルの各ラベルに，o との関連の強さを表すランクを与える．o に強く関連するラベルを主ラベル（Primary Label），主ラベルほどではないが o に関連するラベルを副ラベル（Secondary Label）とし，それぞれの集合を $P(o)$，$S(o)$ で表す．$L(o)$ のラベルは主ラベルもしくは副ラベルなので $L(o) = P(o) \cup S(o)$ かつ $P(o) \cap S(o) = \phi$ である．また，主ラベルは o と最も強く関連しているラベルで，少なくとも1つは存在するため $P(o) \neq \phi$ である．

Series of Criteria for Data with Ranked Multi-Labels
† Masahiro Kuzunishi, Faculty of Economics, Aichi Gakuin University
‡ Tetsuya Furukawa, Faculty of Economics, Kyushu University

3 関連の強さに対する評価基準

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 $\mathcal{L}$ への関連が強いことを $o_1 <_{\mathcal{L}} o_2$，$\mathcal{L}$ が明らかなときは $o_1 < o_2$ で表す．o_2 は条件 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 $\mathcal{L}$ へ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 との関連の強さには，(1)o は $\mathcal{L}$ に関連するか（o は $\mathcal{L}$ のいずれかのラベルに関連するか），(2)o は $\mathcal{L}$ のすべてのラベルに関連するか，(3)o は $\mathcal{L}$ と無関係なラベル L（$L \not\preceq L', L' \not\preceq L, \forall L' \in \mathcal{L}$）に関連しないかの3つの観点があり，それらは次の評価基準となる．

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

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

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

これらを主ラベルの性質としても評価基準となる．

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

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

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

$\mathcal{L}$ との関連は，副ラベルよりも主ラベルが優先されるので，副ラベルの性質は評価基準とはならない．

評価基準間の強さの関係が分かれば，オブジェクトのラベル集合への関連の強さを段階的かつ一元的に評価できるようになる．評価基準間の強さの関係は，評価基準の含意で判断できる [2]．

［定義 1］ 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 $\mathcal{L}$ の評価基準であるといい $cnd_1 < cnd_2$ で表す． □

ラベル集合 $\mathcal{L}$ の $\overline{\mathcal{L}}$ において，cnd を満足する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 $\overline{\mathcal{L}} = \overline{\mathcal{L}}^{LE}$ である．

評価基準間の強さの関係は，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．　□

評価基準を満たすオブジェクト集合の包含関係は，評価基準に含意があるどうかで判断できる．すなわち，ラベル集合 $\mathcal{L}$ と評価基準 cnd_1，cnd_2 に対し，$\overline{\mathcal{L}}^{cnd_2} \subseteq \overline{\mathcal{L}}^{cnd_1}$ と $cnd_2 \Rightarrow cnd_1$ は等価である．補題 1 より評価基準間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ため，評価基準間の強さの関係は評価基準の含意で判断できる．

［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．　□

4　評価基準系列

PE，PA，PN は，主ラベルを対象にそれぞれ LE，LA，LN を満たすかというもので，それらの評価基準と含意があり強さを比較できる．

［補題 2］　$LE < PE$，$LA < PA$，$PN < LN$ である．　□

PN を満たすオブジェクト o の主ラベルは，ラベル集合 $\mathcal{L}$ と無関係なラベルに関連しないがランク付集合ラベルの性質より $P(o) \neq \phi$ なので $\mathcal{L}$ に関連する．よって，o は PE を満たすため，2 つの評価基準には含意があり強さを比較できる．

［補題 3］　$PE < PN$ である．　□

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る．

［補題 4］　$LE < LA$，$PE < PA$ である．　□

評価基準 cnd_x と cnd_y（$cnd_x, cnd_y \in \{LE, PE, LA, PA, LN, PN\}, cnd_x \neq cnd_y$）の両方を条件とする評価基準を $cnd_x \cdot cnd_y$ とする．補題 2，3，4 とそれらの推移律で得られる評価基準間の強さから，cnd_x と cnd_y が互いに含意しない組合せで $LA \cdot PE$，$LA \cdot PN$，$LA \cdot LN$，$PA \cdot PN$，$PA \cdot LN$ が新たな評価基準となる．また，$cnd_x < cnd_x \cdot cnd_y$（$cnd_x$ は cnd_y を含意しないとき），$cnd_x \cdot cnd_y < cnd_x \cdot cnd_{y'}$（$cnd_y < cnd_{y'}$ のとき）である．

PA は $LA \cdot PE$ を含意するので，評価基準間の強さの関係は図 1 となる．矢印の方向は強い評価基準を示す．

図 1 における LE から $PA \cdot LN$ までの経路は，次に示す評価基準間の強さの順に該当する (LE と $PA \cdot LN$ は省略)．

i) $LA < LA \cdot PE < PA < PA \cdot PN$
ii) $LA < LA \cdot PE < LA \cdot PN < PA \cdot PN$
iii) $LA < LA \cdot PE < LA \cdot PN < LA \cdot LN$
iv) $PE < LA \cdot PE < PA < PA \cdot PN$

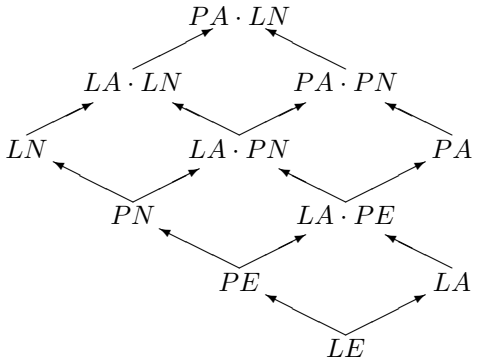

図 1: 評価基準間の強さの関係

v) $PE < LA \cdot PE < LA \cdot PN < PA \cdot PN$
vi) $PE < LA \cdot PE < LA \cdot PN < LA \cdot LN$
vii) $PE < PN < LA \cdot PN < PA \cdot PN$
viii) $PE < PN < LA \cdot PN < LA \cdot LN$
ix) $PE < PN < LN < LA \cdot LN$

1 つの経路に含まれる評価基準の部分集合を評価基準系列あるいは単に系列と呼ぶ．系列は全順序集合なので，ラベル集合に対するオブジェクトの関連の強さを評価する 1 つの体系となる．すなわち，オブジェクトが系列のどの評価基準まで満たすかによって関連の強さが段階的に決まり，オブジェクト間で強さを一元的に比較できる．例えば，i) の系列では，オブジェクトがラベル集合 $\mathcal{L}$ のすべてに関連するという評価基準が優先され，次に $\mathcal{L}$ 内のラベルに強く関連するかで評価し，どの評価基準までを満たすかでオブジェクトの $\mathcal{L}$ への関連の強さが決まる．LA を満たすオブジェクトは満たさないものよりも $\mathcal{L}$ への関連が強く，$LA \cdot PE$ を満たすものはさらに関連が強いといったような比較ができる．

5　おわりに

データの検索や分析の用途に応じて評価基準系列を選択することで，データのラベル集合へ関連の強さを段階的かつ一元的に評価できる．データがラベル集合 $\mathcal{L}$ のすべてに関連することを優先させる，$\mathcal{L}$ 内のラベルに強く関連することを優先させる，$\mathcal{L}$ と無関係なラベルに関連しないことを優先させるなどの系列やこれらを交互に優先させる系列などを用いることで，検索や分析における多様な用途に対応できる．

参考文献

[1] Tang, L., Rajan, S., and Narayanan, V., “Large Scale Multi-label Classification via MetaLabeler,” *Proc. Int’l Conf. on World Wide Web* (WWW ’09), pp. 211–220, 2009.

[2] Kuzunishi, M. and Furukawa, T., “Strength of Relationship Between Multi-labeled Data and Labels,” *Lecture Notes in Computer Science*, Vol. 9357, Springer, pp. 99–108, 2015.