地域分析
第58巻第1・2合併号
2019年12月

一般論文

ランク付集合ラベルデータの分析手法と人工データによる有用性の検証

葛西　正裕

目次:

概要

多様なデータに対する分析手法が求められている．データの意味をラベルで表すデータモデルでは，複数のラベル（集合ラベル）がデータに付されることが多い．データと集合ラベルの各ラベルとの関連の強さをランクとして与えることにより，精緻な分析が可能になる. 分析対象を表すラベル集合とデータの集合ラベルとの関連の強さを評価する基準を導入することで，分析対象に対するデータの関連度は定性的に決まり，その差異を利用した分析が可能になる．本論文は，評価基準によって定まるデータの関連度に基づいた分析手法を実装し，人工的に生成したランク付集合ラベルデータに適用することで，分析手法の有用性を検証する．関連度を段階的に高めてデータを集約した結果，データ件数の変化は，分析対象の重要度，集中度，閉鎖度を表す指標として有用であることを確認した．また，量的データの集約値の変化を評価基準の違いで解釈することで，分析対象に関する有用な知見が得られることも確認した.

キーワード: ランク付集合ラベル，人工データ，分析手法，データモデル，Python

Verification of the Usefulness of an Analysis Method by Synthetic Ranked Multi-Label Data

Masahiro KUZUNISHI

Contents:

Abstract

The importance of analysis methods for heterogeneous dataset is increasing. In the data model where data semantics are expressed as labels, individual data of such dataset is usually annotated with multi-label or a set label. When ranks are given to labels of a set label to express the strength of relationship between data and the set label, data can be analyzed more precisely. Criteria for the strength of relationship between data and a set of labels can evaluate the relevance degree of data to an analysis subject qualitatively. The strength difference can be used to data analysis. The purpose of this paper is to verify the usefulness of the analysis method, which is based on the strength of relationship between data and an analysis subject, by analyzing synthetic ranked multi-label data with the implemented analysis method. As the result of aggregations of data satisfying each of criteria, the change of the number of data becomes to be an index to measure the degree of importance, concentration and closure of the analysis subject. Moreover, the change of aggregate values of the quantitative attribute can be explained by the difference of criteria, and it gives useful information on the analysis subject.

Keywords: Ranked Multi-Label, Synthetic Data, Analysis Method, Data Model, Python

I. はじめに

今日において収集されるデータは，従来の数値データに加え，テキスト，画像，音声，動画データといったように多様化している[HUWCL14]．多様なデータの種類を問わず分析するには，データを適切に構成しておく必要があり，データの意味に応じてラベルを付しておくことが有用である．

1 件のデータには，データの意味に応じて複数のラベル（集合ラベル）が付されることが多い[BIK11] [LOPOW12] [WUWZYLZZ15]．集合ラベルが付されたデータの分析に対し，論文[FUK10]では，様々な集約範囲があることを明らかにしており，目的に応じた集約によって多面的な分析を可能にする理論的枠組みが示されている．しかし，先行研究は，データと集合ラベルの各ラベルとの関連の強さについて言及しておらず，ラベルに対するデータの関連度を分析に反映できない．

データとの関連の強さをランクとしてラベルに与えることで，より精緻な分析が可能になる．例えば，ある企業が複数の地域に展開している場合，展開地域の集合ラベルの各ラベルに地域ごとの営業利益率や売上高といった経営指標，本社機能や研究開発機能の有無などに基づいてランクを付すことで，展開の程度に応じたデータの集約ができる．具体的に述べると，米国や中国に主に展開している企業群と米国や中国にも展開しているが他国が主である企業群を営業利益率の平均値に基づいて比較するといったことが可能になる．

ランク付集合ラベルが付されていることで精緻な分析が可能になるが，分析対象をラベル集合で与えてデータを集約する際，どのような範囲でデータを集約するかが問題になる．論文[KUF15]は，データとラベル集合との関連の強さを評価する基準を導いたうえで，評価基準の強さの順序を明らかにしている．これにより，データがどの評価基準まで満たすかで分析対象に対するデータの関連度を定性的に定めることができる．関連度の程度に応じたデータの集約が可能になり，関連度の差異を利用した分析ができる．例えば，{米国，中国}といったラベル集合に強く関連している企業群ほど，営業利益率の平均値が高くなるといった知見が得られる．

論文[KUF15]は，ランク付集合ラベルデータに対する分析手法を提案しているが，理論的枠組みにとどまっているため，本論文は，提案手法の有用性を実験的に検証する．実験的な検証では，人工データを用いて提案手法を評価することが行われている[SUS14] [TAIKS14]．ランク付集合ラベルを持つ実データを取得することは容易ではないため，ランク付集合ラベルが付されたデータを人工的に生成したうえで，分析手法を適用し，その有用性を検証する．

本論文の構成は以下の通りである．第 II 節は，データと集合ラベルの各ラベルとの関連の強さを表すランクを導入し，データと分析対象を表すラベル集合との関連の強さを評価する基準を示す．第 III 節は，評価基準の強さの順序を明らかにした上で，データと分析対象との関連の強さの差異に基づいた分析手法の理論的枠組みを紹介する．第 IV

節は，ランク付集合ラベルデータを人工的に生成する方法を述べる．第 V 節は，分析手法を実装し，人工データに適用した結果を示しながらその有用性を検証する．第 VI 節は本論文のまとめである．

II.　ランク付集合ラベルと評価基準

分析に供されるデータセットにおける 1 件のデータには，概念階層における複数のカテゴリ（クラスとも呼ばれる[REPLDR14]）のラベル（タグとも呼ばれる[STDMLT11]）が付されていることが多い[BIK11] [LOPOW12] [WUWZYLZZ15]．ラベル集合はデータの意味を表している．例えば，企業群で構成されるデータセットの 1 件のデータにおいて，企業が 3 カ国に展開していることを{日本，中国，米国}といったラベル集合で表している．地域という属性以外にも，輸送機器，電子部品など業種に関するラベル集合も付されることがある．論文[KU-10]は，複数の属性を持つデータに関する記述法を述べているが，本論文では，議論を簡単にするため単一属性のデータを想定する．

ラベル集合の要素間には一般に属性ごとに階層的な関係があり[BIK11] [REPLDR14]，例えば，地域には，日本，東海地方，愛知，名古屋といった階層構造がある．本論文では，データには階層における最下層のカテゴリのラベルが付されているものとし，階層構造を意識した分析が行われるものとする．

データセットの 1 件のデータを d，ラベルを l，ラベル集合を $L = \{l_1, \cdots, l_n\}$とする．dに付されたラベル集合を集合ラベルと呼び，$L(d)$ $(\neq \varphi)$で表す．ラベルの概念の上下関係を>で表す．ラベル l_1 と l_2 に対し，l_2 が l_1 の上位概念のラベルならば，l_2 は l_1 の上位であり，$l\_2 > l\_1$ となる．l_2 が l_1 の上位または l_2 と l_1 が等しいとき，$l\_2 \gtrsim l\_1$ で表し，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_1 中のラベルでラベル集合 L_2 に関連するラベルの集合を $Rel_L_2(L_1) = \{l \mid l \in L_1, \exists l' \in L_2, l' \gtrsim l\}$で表す．データとラベル集合との関連の強さの評価に用いるラベル集合を $\boldsymbol{L}$ とし，$L(d)$中のラベルが $\boldsymbol{L}$ に関連する，すなわち，$Rel\_\boldsymbol{L}(L(d)) \neq \varphi$ ならば，d は $\boldsymbol{L}$ に関連する．

$L(d)$の各ラベルに，d との関連の強さを表すランクを与える．d に強く関連する $L(d)$ 中のラベルを主ラベル（Primary Label），主ラベルほどではないが d に関連する $L(d)$中のラベルを副ラベル（Secondary Label）という．$L(d)$中の主ラベルの集合を $P(d)$，副ラベルの集合を $S(d)$で表し，集合ラベルはランク付集合ラベルとする．$L(d)$のラベルで d に最も強く関連しているラベルは主ラベル，$L(d)$のラベルは主ラベルもしくは副ラベルと考えられるので，ランク付集合ラベルには以下の性質がある．

[性質 1] データ d の $L(d)$には主ラベルが必ず含まれる:$P(d) \neq \varphi$

[性質 2] データ d の $L(d)$は主ラベルと副ラベルの集合からなる:$L(d) = P(d) \cup S(d)$

[性質 3] データ d の $L(d)$のラベルは主ラベルか副ラベルである:$P(d) \cap S(d) = \varphi$

分析対象をラベル集合 $\boldsymbol{L}$ で与え，データと $\boldsymbol{L}$ との関連の強さに応じて集約することで，関連の強さに基づいて集約値を評価するといった分析が可能になる．これより，データと $\boldsymbol{L}$ との関連の強さを評価する基準を示す．データ d_1 と d_2 に対し，d_2 の方が d_1 よりもラベル集合 $\boldsymbol{L}$ との関連が強いことを $d\_2 >\_{\boldsymbol{L}} d\_1$，$\boldsymbol{L}$ が明確であれば $d\_2 > d\_1$ で表す．データ d_1 と d_2 及び条件 *cnd* に対し，d_2 は *cnd* を満たし d_1 は *cnd* を満たさないとき $d\_2 > d\_1$ ならば，*cnd* はデータと $\boldsymbol{L}$ との関連の強さの評価基準とする．

データ d とラベル集合 $\boldsymbol{L}$ に対し，

LE: d は $\boldsymbol{L}$ に関連する（$Rel\_\boldsymbol{L}(L(d)) \neq \varphi$），

PE: d は $\boldsymbol{L}$ に強く関連する（$Rel\_\boldsymbol{L}(P(d)) \neq \varphi$），

LN: d が関連するのは $\boldsymbol{L}$ のみである，言い換えると，

d は $\boldsymbol{L}$ 以外に全く関連しない（$Rel\_\boldsymbol{L}(L(d)) = L(d)$），

PN: d が強く関連するのは $\boldsymbol{L}$ のみである，言い換えると，

d は $\boldsymbol{L}$ 以外に強く関連しない（$Rel\_\boldsymbol{L}(P(d)) = P(d)$）

という条件が評価基準である．

上記以外の評価基準には，ラベル集合 $\boldsymbol{L}$ に関連するラベルの数が多いデータほど $\boldsymbol{L}$ に強く関連するという条件も考えうるが，例えば，ラベル集合 $\boldsymbol{L}$ = {輸送機器}に対し，四輪自動車のみを展開する企業の方が，四輪自動車と二輪自動車に加え金融も展開している企業よりも $\boldsymbol{L}$ に強く関連していると考えられる．また，対象領域によっても関連するラベルの数と関連の強さとの関係には様々な考え方がある．よって，本論文では，関連するラベルの数は考慮しないものとする．

III. 分析手法の理論的枠組み

評価基準の強さの順序が分かれば，データとラベル集合との関連の強さを一元的に評価できる．本節は，論文[KUF15]をもとに，データとラベル集合との関連の強さの差異に基づく分析手法の理論的枠組みについて紹介する．

評価基準を満たすデータセットと満たさないデータセットの集約値を比較することで分析ができる．米国と中国の関係に着目した企業データの分析を例にあげる．分析対象をラベル集合 $\boldsymbol{L}$ = {米国，中国}で与え，*PE* を満たすデータセットと満たさないデータセットの営業利益率の平均値を比較することで，米国や中国に主に展開する企業群の方がそうではない企業群よりも営業利益率が高いといった知見が得られたりする．このように評価基準は単独で分析に用いることができる．一方，評価基準の強さの順序は評価基準の含意で判断できるので，評価基準の強さの順序が明らかになれば，データとラベル集合との関連の強さを一元的に評価できる．例えば，*PE* を満たすデータセットのうち *PE* よりも強い評価基準である *PN* を満たすデータセット，すわなち，米国や中国のみに主に展開している企業群で集約できる．そのような企業群の営業利益率の平均値がさらに高ければ，米国や中国に集中した企業群がより高い利益を上げていることが分かる．

ラベル集合 $\boldsymbol{L}$ と評価基準 C_1 と C_2 に対し，C_1 を満たすデータ d_1 が C_2 を満たさず，データ d_2 が C_2 を満たすとき，$d\_2 > d\_1$ ならば，C_2 は C_1 よりも強い $\boldsymbol{L}$ の評価基準であるといい，$\boldsymbol{L}$ が明確なとき $C\_2 > C\_1$ で表す．

評価基準の強さの順序は評価基準の含意で判断できる．すなわち，評価基準 C_1 と C_2 に対し，$C\_2 > C\_1$ と $C\_2 \Rightarrow C\_1$ は等価である．評価基準の強さの順序について，以下の補題が成り立つ．

【補題 1】 データ d とラベル集合 $\boldsymbol{L}$ に対し，$PE > LE$ である． □

【証明】 d が PE を満たせば $P(d)$ に $\boldsymbol{L}$ に関連するラベルが存在するので，d は LE も満たす．よって，$PE \Rightarrow LE$ であり $PE > LE$ である． □

【補題 2】 データ d とラベル集合 $\boldsymbol{L}$ に対し，$PN > PE$ である． □

【証明】 d が PN を満たせば $P(d)$ のラベルはすべて $\boldsymbol{L}$ に関連する．性質 1 より $L(d)$ には主ラベルが必ず含まれるので，$P(d)$ には $\boldsymbol{L}$ に関連するラベルが存在するため，d は PE も満たす．すなわち，$PN \Rightarrow PE$ であり $PN > PE$ である． □

【補題 3】 データ d とラベル集合 $\boldsymbol{L}$ に対し，$LN > PN$ である． □

【証明】 d は LN を満たせば $L(d)$ のラベルはすべて $\boldsymbol{L}$ に関連する．性質 2 より $L(d) = P(d) \cup S(d)$ なので PN も満たす．すなわち，$LN \Rightarrow PN$ であり $LN > PN$ である．
□

評価基準の強さの順序は定理 1 にまとめられる．

【定理 1】 データ d とラベル集合 $\boldsymbol{L}$ に対し，$LN > PN > PE > LE$ である． □

【証明】補題 1 より $PE > LE$, 補題 2 より $PN > PE$, 補題 3 より $LN > PN$ なので，$LN > PN > PE > LE$ である． □

定理 1 より評価基準の強さの順序が明らかになったので，データ d がどの評価基準まで満たすかによって，ラベル集合 $\boldsymbol{L}$ への関連の強さが定性的に決まる．また，評価基準の強さの順序は評価基準の含意関係と等価であり，含意関係は評価基準を満たすデータ集合の包含関係と等価なので，強さの順序がある評価基準を満たすデータ集合間には包含関係がある．

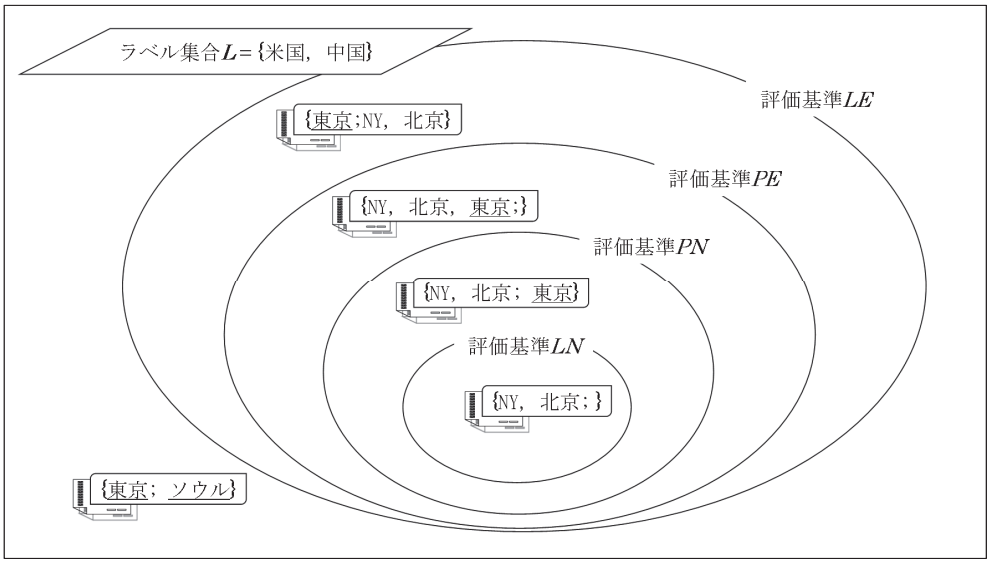
図表 1 は，分析対象をラベル集合 $\boldsymbol{L}$ = {米国, 中国}で与えた際，各評価基準に応じた代表的なデータと包含関係を示したものである．ランク付集合ラベルを主ラベル集合，副ラベル集合の順で表し，$\boldsymbol{L}$ に関連しないラベルには下線を引いている．

{東京; ソウル}のデータは，東京が主ラベル，ソウルが副ラベルであり，どちらも $\boldsymbol{L}$ に関連しないラベルである．よって，いずれの評価基準も満たさず，分析対象に該当しない．{東京; NY, 北京}のデータは LE のみを満たすので，$\boldsymbol{L}$ への関連が最も弱い．次に，{NY, 北京, 東京;}のデータは，$\boldsymbol{L}$ に関連する主ラベル``NY"や``北京"があるので PE まで満たすデータである．{NY, 北京;東京}のデータは，$\boldsymbol{L}$ に関連しない``東京"が副ラベルなので

*PN*まで満たすデータである．そして，{NY, 北京;}のデータは，***L***に関連しないラベルがないので*LN*まで満たし，***L***への関連が最も強い．

評価基準の強さの順序を用いて，*LE*, *PE*, *PN*, *LN*の順序で評価基準を満たすデータ集合を集約することで，***L***に関連する，強く関連する，さらに***L***以外に強く関連しない，全く関連しないという順序で分析対象に対するデータの関連度を段階的に高めて集約できる．

図表 1. 分析対象（ラベル集合）に対する関連度に基づくデータの集約



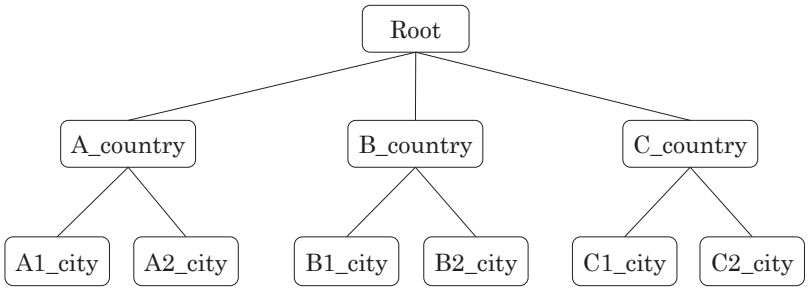
出所：論文[KUF15]をもとに筆者作成

Ⅳ. ランク付集合ラベルを持つ人工データの生成

本節は，ランク付集合ラベルデータを人工的に生成する方法を述べる．まず，ランク付集合ラベルの前提条件を示したうえで，ランク付集合ラベルが付されたデータセットを生成する工程を説明する．次に，ランク付集合ラベルデータセットを集約する際，データ件数に限らず平均値や中央値でも分析できるように定量的なドメインを持つ属性，量的データカラムを追加する．

ランク付集合ラベルデータの前提条件は以下のとおりである． 1 件のデータを 1 企業と考え，100 万件のデータで構成されるデータセットを生成する．企業が展開する地域とその程度をランク付集合ラベルが表すものとし，ラベル集合間には図表 2 の階層構造があるものとする．

図表 2. 人工データにおけるラベル集合の階層構造



出所：筆者作成

データには最下層のカテゴリのラベルが付される．6 都市から構成されるラベル集合を $Cities$ = {A1_city, A2_city, B1_city, B2_city, C1_city, C2_city}とすると，データ d のランク付集合ラベルは，$L(d) = P(d) \cup S(d)$，$P(d) = \{ l \mid l \in Cities \}$，$P(d) \neq \varphi$，$S(d) = \{ l \mid l \in Cities \}$，$P(d) \cap S(d) = \varphi$ を満たす．実装の際は，6 都市のカラムを用意し，主ラベルの状態を 2，副ラベルの状態を 1，ラベルが付されない，すなわちその都市に分類されない状態を 0 で示し，それらをドメインにする．例えば，6 都市のカラムの組(A1_city, A2_city, B1_city, B2_city, C1_city, C2_city)に対し，データ $d_{\_n}$ が(2,0,2,0,0,1)であれば，A1_city の主ラベル，B1_city の主ラベル，C2_city の副ラベルで構成されるランク付集合ラベルが付されていることを示し，n 番目の企業は A1_city と B1_city に主に展開する一方，C2_city にも主ほどではないが展開していると解釈する．

図表 3 は，ランク付集合ラベルデータセットを生成するアルゴリズムである．アルゴリズムを実装した Python[1]のコードは，付録 A-1 の通りである．まず，Step1 では，企業が各都市に展開する状況は独立に生じるものとし，主に展開している確率（主ラベルが付される確率）は 0%から 25%，主といえるほどではないが展開している確率（副ラベルが付される確率）を 0%から 50%の範囲に設定し，その範囲から無作為に値を選択し，都市別ランク確率リストを生成する．例えば，A1_city において， 主ラベル確率リストから 10%，副ラベル確率リストから 20%が選択されると，残り 70%はラベルが付されない確率となり，A1_city のランク確率リストの値が決まる．他の都市についても同様の処理を行い，6 都市のランク確率リストを格納した都市別ランク確率リストを生成する．

Step2 では，都市別ランク確率リストに基づいて，2，1，0 の中から要素を 1 つ選択することでランク付ラベルが決まる．この際，1 件のデータで A1_city から C2_city までのすべてのカラムの値が 0，すなわち主ラベルがないデータが生成されることがある．このようなデータは前提条件に反するため，後から除外することになるが，除外後にデータ件数を確保できるように想定するデータ件数の 2 倍のデータを予め生成する．

Step3 では，主ラベルが 1 つ以上あるデータで構成されるデータセットを生成するた

[1]Python のバージョンは，3.7.3 である．

め，200 万件のデータのうち主ラベルを表す 2 の値が A1_city から C2_city までのいずれかのカラムで存在するデータを選択する．選択されたデータ件数が 100 万件に達しない場合には，Step1 に戻ってランク付集合ラベルを振りなおし，100 万件以上のデータが生成されるまで繰り返す．このようにして生成したデータセットから無作為に 100 万件のデータを抽出し，ランク付集合ラベルデータセットが完成する．

図表 3. ランク付集合ラベルデータセット生成アルゴリズム

アルゴリズム 1 ランク付集合ラベルデータセットの生成

Input: データ件数:*Number* = 1,000,000
都市リスト:*CitesList* = [A1_city, A2_city, B1_city, B2_city, C1_city, C2_city]
ランクリスト:*RankList* = [2:主ラベル, 1:副ラベル, 0:ラベルなし]
主ラベル確率リスト:*Primary* = [0%,･･･, 25%]
副ラベル確率リスト:*Secondary* = [0%,･･･, 50%]
Output: ランク付集合ラベルデータセット:*dataset_labels*
Step1: **都市別ランク確率リストの生成**
1: **for** i **in** *CitesList* **do**
2: *主ラベル確率*_i = *Primary* より無作為抽出
*副ラベル確率*_i = *Secondary* より無作為抽出
3: *ランク確率リスト*_i = [*主ラベル確率*_i, *副ラベル確率*_i,
100% － *主ラベル確率*_i － *副ラベル確率*_i]
4: *都市別ランク確率リスト* =
[[*ランク確率リスト*_A1_city],･･･, [*ランク確率リスト*_C2_city]]
Step2: **全都市のランク付集合ラベルリストの生成**
5: **for** i **in** *CitesList* **do**
6: **for** j **from** 1 **to** *Number* (=1,000,000) * 2 **do**
7: 都市 i の j 番目の値:*l*_i [j] =
*ランク確率リスト*_i のウェイトで *RankList* から 1 つ選択
8: 都市 i のランク付集合ラベルリスト:*Labels*_i =
[*l*_i [1], ･･･, *l*_i [2,000,000]]
9: *全都市のランク付集合ラベルリスト* = [*Labels*_A1_city,･･･, *Labels*_C2_city]
Step3: **主ラベルが存在するデータで構成されるデータセットの生成**
10: *dataset* = {'A1_city ':*Labels*_A1_city,･･･, 'C2_city ':*Labels*_C2_city}
11: *dataset* ← **select** * **from** *dataset* **where** *Labels*_A1_city == 2 **OR**,･･･,
OR *Labels*_C2_city == 2
12: **if number** (*dataset*) < 1,000,000 **then**
Step1 へ戻る
else

dataset_labels ← *dataset* より 1,000,000 個のデータを無作為抽出

出所：筆者作成

ランク付集合ラベルデータセットを集約する際，平均値や中央値等が算出できるようするため量的データカラムを生成し，ランク付集合ラベルデータセットに結合する．図表 4 は量的データカラムを追加するアルゴリズムであり，付録 A-2 はアルゴリズムを実装した Python のコードである．

都市が属する国ごとに経済状況が異なるものとし，良い，普通，悪いという 3 パターンとする．量的データカラムは企業の営業利益率や株価収益率といった指標を想定し，良い，普通，悪い状況での平均値は，それぞれ 10，0，-10 を割り当てる．国ごとに 3 つ値から無作為抽出し，その国の平均値とする．標準偏差は一律に 5 として，正規分布に基づき都市ごとに 100 万件の乱数を発生させ，6 都市分の乱数を 2 次元の表（1,000,000×6）に格納する．例えば，A_country が普通，B_country が悪い，C_country が良い状況（0,-10,10）が設定されたとすると，都市ごとにその都市が属する国の平均値と標準偏差を 5 とする正規分布に基づいた乱数を 100 万個発生させ 2 次元の表に格納し，各都市の経済状況を設定する際に用いる．

企業が展開する都市の経済状況は営業利益率や株価収益率といった指標に影響を与えると考え，展開する都市及びその都市との関連の程度を反映させる．データ d_{n} の量的データカラムの n 番目の値は，2 次元の表における n 番目の値の組を，d_{n} のランク付集合ラベルをウェイトにして加重平均した値とする．例えば，2 次元の表における n 番目の値の組が(9,10,-3,4,-4,-3)，d_{n} のランク付集合ラベルが(2,0,0,0,0,1)ならば，加重平均した 5 が d_{n} の量的データカラムの値になる．このようにして生成された量的データカラムをランク付集合ラベルデータセットに結合し，分析に供されるランク付集合ラベルデータセットが完成する．

図表 4．量的データカラム追加アルゴリズム

アルゴリズム 2 量的データカラムの生成と結合

Input: ランク付集合ラベルデータセット:*dataset_labels*
　　平均値リスト:*MeanList* = [10:良い, 0:普通, -10:悪い]
Output: 量的データカラムがあるランク付集合ラベルデータセット:*dataset_original*

```
for i in [A_country, B_country, C_country] do
        Mean_i = MeanList より無作為抽出
for j in [A1_city,・・・, C2_city] of dataset_labels do
        array_j = 正規分布(平均=Mean_i ( j が属する i ), 標準偏差 = 5,
        個数 = number (dataset_labels))に基づく 1 次元配列
random = {'A1_random ':array_A1_city,・・・, 'C2_random ':array_C2_city}
#random は乱数を格納した 2 次元の表（1,000,000×6）
numerical_value_column = 初期化した 1 次元配列（1,000,000）
```

```
6:    for k from 1 to 1,000,000 do
7:          numerical_value_column [k] =
            加重平均（random [k]，ウェイト= dataset_labels [k]）
8:    dataset_original ← Insert numerical_value_column to dataset_labels
      as 'numerical_value '
```

出所：筆者作成

Ⅴ. ランク付集合データ分析と分析手法の検証

本節は，分析手法を実装し，人工データに適用する．その上で，分析結果を示しながら分析手法の有用性を検証する．

本論文は，図表 2 で示した階層構造を意識した分析を行うことを前提にしているので，分析対象を表すラベル集合 $\boldsymbol{L}$ は国レベルの粒度とし，$\boldsymbol{L}$ = {A_county, B_country}とする．都市レベルの粒度のデータセットを国レベルの粒度の $\boldsymbol{L}$ で集約する際，$\boldsymbol{L}$ の各要素に属するすべての都市レベルのカテゴリのラベルを参照する必要がある．本論文の場合，A1_city から C2_city までのカラムの値を参照する必要がある．本論文が想定するラベルの階層構造は 3 つの国と 6 つの都市からなる深さが 2 といった単純な構造をしており，評価基準を満たすデータを選択するコストは大きくない[2]．一方，予め粒度を固定した分析を行う際には，その粒度のラベルに変換しておくことで，ラベル集合 $\boldsymbol{L}$ の要素数を n，n に属する最下層のカテゴリのラベル集合の要素数を m とすると，計算コストは n/m になる．例えば，日本，米国はそれぞれ 47 都道府県，51 州から構成されるので，都市レベルの粒度のデータセットを国レベルで分析する際に参照するラベルの個数は約 50 分の 1（2/98）になる．このような変換は，同じ粒度でラベル集合の要素を頻繁に入れ替えて集約する際に効率的に処理でき有用である．

分析に供するデータセットに対し，分析対象の粒度のレベルに変換する．データがある国のいずれかの都市に対し，強く関連すればその国と強く関連し，それほどではないが関連すればその国と関連し，関連がなければその国と関連しないと考えられる．よって，都市レベルの粒度のランク付集合ラベルを国レベルの粒度に変換するには，その国に属する都市のラベルのうちで最も高いランクのラベルに置き換えればよい．例えば，6 都市のカラムの組(A1_city, A2_city, B1_city, B2_city, C1_city, C2_city)に対し，データ $d_{\_n}$ が(2, 1, 1, 0, 0, 0)ならば，A1_city の主ラベル，A2_city の副ラベル，B1_city の副ラベルなので，国レベルのカラムの組(A_country, B_country, C_country)に対し，(2, 1, 0)と変換すればよい．図表 4 はラベルの階層粒度変換を行うアルゴリズムを，付録 A-4 はアルゴリズムを実装した Python のコードを示している．

2 実装に用いた Python3.7.3 には，データの前処理に多用されるライブラリの Pandas において階層的なデータを効率的に処理する階層インデックスが用意されており，それを用いて階層構造を持つデータセットの集約を効率的に行うことは可能である．

図表 5. ラベルの階層粒度変換アルゴリズム

アルゴリズム 3 ラベル階層粒度変換

Input: *dataset_original*
階層構造の定義:A_country = ['A1_city', 'A2_city'],
B_country = ['B1_city', 'B2_city'], C_country = ['C1_city', 'C2_city']

Output: 国レベルの粒度のラベルに変換されたランク付集合ラベルデータセット: *dataset_transformation*

1: *A_country_array* = 初期化した 1 次元配列（1,000,000）

2: **for** i **from** 1 **to number** (*dataset_original*) **do**

3: *dataset_original* のカラム A1_city と A2_city の i 番目の値に対し,
A_county_array [i] = **Max** ({A1_city[i], A2_city[i]})

4: *B_country_array*, *C_country_array* についても同様

5: *dataset_transformation* = {'A_country':*A_country_array*,
'B_country':*B_country_array*, 'C_county':*C_country_array*,
dataset_original [numerical_value]}が完成

出所：筆者作成

分析に供されるデータセット $\boldsymbol{D}$ とする．ラベル集合 $\boldsymbol{L}$ に対し，評価基準 *cnd*（*cnd* ∈ {*LE*, *PE*, *PN*, *LN*}）を満たす $\boldsymbol{D}$ のデータの集合を $\boldsymbol{D}(\boldsymbol{L}, cnd)$で表す．分析対象をラベル集合 $\boldsymbol{L}$ = {A_county, B_country}で与え，評価基準 *LE*, *PE*, *PN*, *LN* の 4 段階で $\boldsymbol{L}$ との強さを段階的に強めて集約を行う．分析手法を実装した Python コードは付録 A-4 の通りである．

まず，*LE* を満たすデータは主ラベルまたは副ラベルが $\boldsymbol{L}$ に関連していればよいので，$\boldsymbol{D}(\boldsymbol{L}, LE)$は，国レベルのラベルに変換されたデータセット *dataset_transformation* における A_country または B_country の値が 2 もしくは 1 のデータである．次に，*PE* を満たすデータは主ラベルが $\boldsymbol{L}$ に関連していればよいので，$\boldsymbol{D}(\boldsymbol{L}, PE)$は，A_country または B-_country の値が 2 のデータである．さらに，*PN* を満たすデータは $\boldsymbol{L}$ 以外に関連する主ラベルがなければよいので，$\boldsymbol{D}(\boldsymbol{L}, PN)$は，C_county の値が 2 ではない，すなわち 0 もしくは 1 のデータである．[性質 1]より，このようなデータは A_country または B_country の値が必ず 2 になるので，A_country や B_country の値を設定する必要はない．最後に，*LN* を満たすデータは $\boldsymbol{L}$ 以外に関連する主ラベルや副ラベルがなければよいので，$\boldsymbol{D}(\boldsymbol{L}, LN)$は，C_county の値が 0 のデータが該当し， A_country や B_country の値は同様に設定する必要はない．

各評価基準を満たすデータ件数の変化から分析対象に関する有用な情報が得られる．第一に，$\boldsymbol{D}$ 中の $\boldsymbol{L}$ に関連するデータの集合に対し，L に強く関連するデータの集合の割合が高ければ，A_country や B_country に強く関連する傾向があり，それらの国の重要性が分かる．第二に，$\boldsymbol{D}$ 中の $\boldsymbol{L}$ に強く関連するデータの集合に対し，$\boldsymbol{L}$ 以外に強く関連しないデ

ータの集合の割合が高ければ，C_country には強く関連せず A_country や B_country のみに強く関連する傾向があり，A_country や B_country への集中度が高いことが分かる．第三に，$\boldsymbol{D}$ 中の $\boldsymbol{L}$ 以外に強く関連しないデータの集合に対し，$\boldsymbol{L}$ 以外に全く関連しないデータの集合の割合が高ければ，C_country には全く関連せず A_country や B_country のみに強く関連する傾向があり，A_country や B_country の閉鎖性が高いことが分かる．

まとめると，データセット $\boldsymbol{D}$ を評価基準 LE, PE, PN, LN の順序で $\boldsymbol{L}$ への関連の強さを段階的に強めて集約した際，関連度の差異によって生じる個数の変化は，$\boldsymbol{L}$ の重要度，集中度，閉鎖度と解釈でき，以下の指標が導かれる．

データセット $\boldsymbol{D}$ とラベル集合 $\boldsymbol{L}$ に対し，

重要度指数:$Index\_{Importance}(\boldsymbol{D}, \boldsymbol{L}) = |\boldsymbol{D}(\boldsymbol{L}, PE)| \,/\, |\boldsymbol{D}(\boldsymbol{L}, LE)|$,

集中度指数:$Index\_{Concentration}(\boldsymbol{D}, \boldsymbol{L}) = |\boldsymbol{D}(\boldsymbol{L}, PN)| \,/\, |\boldsymbol{D}(\boldsymbol{L}, PE)|$,

閉鎖度指数:$Index\_{Exclusiveness}(\boldsymbol{D}, \boldsymbol{L}) = |\boldsymbol{D}(\boldsymbol{L}, LN)| \,/\, |\boldsymbol{D}(\boldsymbol{L}, PN)|$

と定義する．なお，いずれの指数も 0 から 1 の範囲の値をとり，1 に近いほど程度が高いと評価できる．

本論文で用いた人工データを生成するアルゴリズムは，設定値を無作為に選んだり乱数を発生させたりしているので生成する度にデータセットは異なる．そこで，人工データを Case1 から Case10 までの 10 通りを生成し分析を行った．結果は図表 6 の通りである．

図表 6. 人工データの分析結果

	LE				*PE*					*PN*					*LN*				
	個数	平均	中央値	標準偏差	個数	重要度指数	平均	中央値	標準偏差	個数	集中度指数	平均	中央値	標準偏差	個数	閉鎖度指数	平均	中央値	標準偏差
Case1	879,113	10.0	10.0	3.9	730,965	0.83	10.0	10.0	4.0	571,129	0.78	10.0	10.0	4.1	478,662	0.84	10.0	10.0	4.3
Case2	822,546	5.4	5.3	3.9	541,849	0.66	6.5	6.3	3.6	317,570	0.59	7.4	7.3	3.7	66,833	0.21	10.0	10.0	4.2
Case3	961,417	0.4	0.4	5.0	718,477	0.75	0.6	0.7	5.3	520,257	0.72	0.6	0.8	5.7	102,466	0.20	0.8	1.0	7.3
Case4	893,888	-1.6	-1.7	6.5	818,228	0.92	-2.3	-2.2	6.2	674,417	0.82	-3.0	-3.1	6.3	194,984	0.29	-7.0	-8.0	7.2
Case5	875,518	-3.2	-3.5	7.3	755,039	0.86	-2.3	-2.3	7.3	580,835	0.77	-1.6	-1.4	7.8	244,000	0.42	0.2	0.8	9.1
Case6	977,454	-0.3	-0.2	5.1	866,817	0.89	-0.9	-0.8	4.9	722,893	0.83	-1.6	-1.5	4.8	467,644	0.65	-2.7	-2.7	4.9
Case7	965,906	-6.7	-6.6	3.7	770,135	0.80	-7.4	-7.3	3.5	593,301	0.77	-7.9	-7.8	3.5	156,015	0.26	-10.0	-10.0	3.8
Case8	964,265	-1.2	-1.1	5.3	848,674	0.88	-1.8	-1.7	5.1	679,792	0.80	-2.7	-2.7	4.9	514,729	0.76	-3.6	-3.5	4.9
Case9	958,132	4.2	4.1	4.2	764,966	0.80	4.9	4.8	4.0	588,019	0.77	5.3	5.2	4.2	215,573	0.37	6.2	6.2	4.8
Case10	975,152	3.4	3.4	3.7	727,697	0.75	2.6	2.7	3.5	503,994	0.69	1.8	1.9	3.4	155,467	0.31	0.0	0.0	3.6

出所：筆者作成

Case4 では，LE を満たす 893,888 件のデータのうち，818,228 件のデータが PE を満たすので，$\boldsymbol{L}$ の重要度指数は 0.92 である．すなわち，A_country または B_country に展開する企業のうち 92%がそれらの国に主に展開していることが分かり，A_country や B_country の重要度が高いことが分かる．一方，Case2 では，$\boldsymbol{L}$ の重要度指数は 0.66 に留まっている．

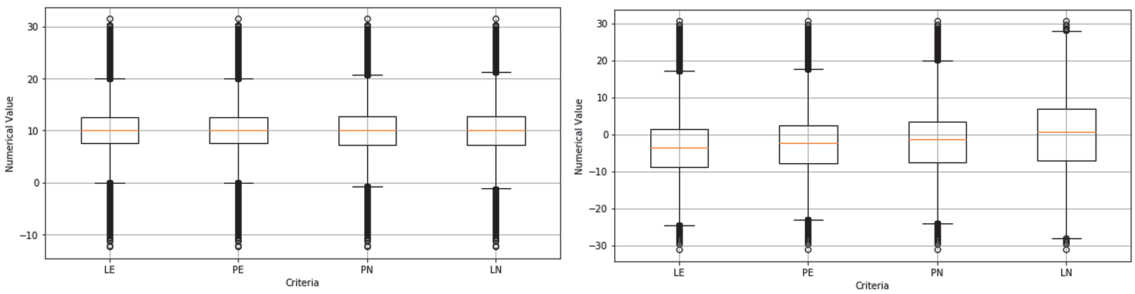
$\boldsymbol{L}$ の集中度指数については，Case6 が最も高い値であり，A_country または B_country に主に展開する企業のうち 83%が C_country には主に展開せず A_country または

B_country のみに主に展開していることから，A_country や B_country への集中度が高いことが分かる．一方，Case2 の集中度指数は 0.59 に留まっており，約 4 割の企業が C_county にも主に展開しているとも解釈できる．

$\boldsymbol{L}$ の閉鎖度指数については，Case1 が最も高い値であり，A_country または B_country のみに主に展開している企業のうち 84%が C_country には全く展開しておらず，A_country や B_country の閉鎖度が高いことが分かる．一方，Case3 の閉鎖度指数は 0.2 に留まっており，8 割の企業が C_county にも主ではないが展開していると解釈できる．

図表 7 は，量的データカラムの集約値の変化が対照的な Case1 と Case5 について，箱ひげ図を用いて示したものである．Case1 の中央値は，$\boldsymbol{L}$ との強さを段階的に強めて集約しても変化が見られない．すなわち，評価基準を変えても中央値が変化していないことから，$\boldsymbol{L}$ に強く関連する，さらに $\boldsymbol{L}$ 以外に強く関連しない，$\boldsymbol{L}$ に全く関連しないという基準が値に影響を与えないことが示唆される．一方，Case5 では，$\boldsymbol{L}$ への関連度が高まるほど，中央値が高くなっている．A_country または B_country に展開している企業のうち，主に展開している企業群の方が営業利益率や株価収益率を想定した値の中央値が高く，展開の程度と企業収益に正の相関があると考えられる．さらに，C_county に主に展開しない企業群，全く展開していない企業群の収益性はより高くなっており，C_county への展開よりも A_country または B_country に重点的に展開する方が良いことが示唆される．図表 6 から，Case2，Case3，Case9 にも同様の傾向がある．反対に，Case4，Case6，Case7，Case8，Case10 のように $\boldsymbol{L}$ への関連の程度が高まるほど企業収益が悪化している場合もあり，A_country または B_country に展開することの負の影響や C_country で展開する有望性などが伺える．

図表 7. 量的データカラムの集約値の変化



Case1　　Case5

出所：筆者作成

VI. おわりに

本論文は，ランク付集合ラベルデータを人工的に生成したうえで，分析対象に対するデータの関連度の差異に基づく分析手法を実装し，その有用性を検証した．データと分析対象との関連の強さを評価する基準は，*LE*, *PE*, *PN*, *LN*であり，分析対象に関連する，強く関連する，さらに分析対象以外に強く関連しない，全く関連しないという順序で関連度を段階的に高めてデータを集約できる．

集約値の変化は，評価基準に違いによって生じるものであり，分析対象に関する有用な知見が得られる．関連度が高まることで生じるデータ件数の変化は，分析対象の重要度，集中度，閉鎖度を表しており，人工データを用いた実験により，それら指標の有用性を確認した．また，生成した人工データによって，関連度が高まるごとに数値データの集約値が高くなる，低くなる，変化がない場合があることを確認した．本論文が想定した事例でいえば，企業の収益性と分析対象への関連度の高さに正の相関があれば，分析対象に強く関連する企業群，さらに分析対象以外に強く関連しない企業群，全く関連しない企業群ほど収益性が高くなるので，分析対象の国の都市に重点的に展開する方が良いのではないかと推測できる．

ランクを有するデータに関する研究は，主に情報検索の分野で行われており，利用者の嗜好を数値化して評価し，高度な検索を実現している[WABMN16] [WAHCSWC13]．既存研究は情報検索を目的としているのに対し，本論文は，データ分析の重要性が増すなかで，ランク付集合ラベルデータに対する分析手法の有用性を確認した．

本論文の分析手法は様々な経済分析に適用できる．例えば，労働形態の多様化が進む中で，複数の企業や異なる業種において，本業と同程度に働く兼業の場合や副業をしている場合が増えている．兼業には主ラベル，副業には副ラベルを与えることで，労働形態と収入や満足度の関係を分析できる．さらに，本論文の分析手法は経済分析のみならずランク付集合ラベルを有するよう様々な領域において応用可能である．

本論文では，分析対象に関連するラベルの個数は問わない評価基準を採用した．一方，分析対象のすべてに関連するデータがそうではないデータよりも強く関連するという考え方ができる．これにより，考慮すべき評価基準は増え，評価基準の強さの順序は複雑になる．そうした評価基準に基づく分析手法については，論文[KUF17]や[KUF18]で議論しているので，その有用性を今後検証したい．また，本論文では，ランクの区分を主ラベルと副ラベルの 2 種類としたが，区分が多いほど精緻な分析が可能になる．論文[KUF18]では，ランクの区分を k 個とするランク付集合ラベルに対する分析のための理論的枠組みを提案しているので，k ランクの集合ラベルデータに対する分析手法の有用性に関しても検証を行いたい．

参考文献

[BIK11] Bi, W. and Kwok, J.:Multi-Label Classification on Tree- and DAG-Structured Hierarchies, *Proc. Int'l Conf. on Machine Learning*, pp.17-24 (2011).

[FUK10] 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).

[HUWCL14] 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).

[KU10] 葛西正裕:多重属性を持つラベル集合を用いたデータの記述, 愛知学院大学産業研究所所報「地域分析」, Vol.49, No.1, pp.43-65 (2010).

[KUF15] 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).

[KUF17] 葛西正裕, 古川哲也:ランク付集合ラベルデータのための評価基準系列, 情報処理学会第 79 回全国大会, 3B-04 (2017).

[KUF18] 葛西正裕, 古川哲也:ランク付集合ラベルデータの分析のための理論的枠組み, 第 10 回データ工学と情報マネジメントに関するフォーラム（第 16 回日本データベース学会年次大会）, I6-4 (2018).

[LOPOW12] Lopez, V., Prieta, F., Ogihara, M., and Wong, D.:A Model for Multi-Label Classification and Ranking of Learning Objects, *Expert Systems with Applications*, Vol.39, Issue 10, pp.8878-8884 (2012).

[REPLDR14] 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*, pp.213-222 (2014).

[STDMLT11] 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*, pp.109-112 (2011).

[SUS14] 数原良彦, 櫻井彰人:順序ラベルを用いた教師あり自己組織化マップ, 知能と情報（日本知能情報ファジィ学会誌）, Vol.26, No.5, pp.809-819 (2014).

[TAIKS14] 竹内孝, 石黒勝彦, 木村昭悟, 澤田宏:非負制約下における複合行列分解とそのソーシャルメディア解析への応用, 情報処理学会論文誌:数理モデルと応用, Vol.7, No.1, pp.71-83 (2014).

[WABMN16] Wang, X., Bendersky, M., Metzler, D., and Najork, M.:Learning to Rank with Selection Bias in Personal Search, *Proc. ACM Int'l Conf. on Research and Development in Information Retrieval,* pp. 115-124 (2016).

[WAHCSWC13] Wang, H., He, X., Chang, M., Song, Y., White, R., and Chu, W.: Personalized Ranking Model Adaptation for Web Search, *Proc. ACM Int'l Conf. on Research and Development in Information Retrieval*, pp. 323-332 (2013).

[WUWZYLZZ15] 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).

付録

A-1 ランク付集合ラベルデータセット生成アルゴリズムのコード

```
import numpy as np
import pandas as pd

DATA_NUMBER = 1000000
CITIES_LIST = ['A1_city', 'A2_city', 'B1_city','B2_city','C1_city', 'C2_city']
RANK_LIST = [2,1,0] #2:主ラベル 1:副ラベル, 0:ラベルなし
PRIMARY_RATIO_LIST = list(range(26)) #主ラベルが付される確率:0%～25%
SECONDARY_RATIO_LIST = list(range(51)) #副ラベルが付される確率:0%～50%

data_counter = 0
city_number = len(CITIES_LIST)
while (data_counter < DATA_NUMBER):

#Step1. 都市別ランク確率リストの生成
    ratio_list =[]
    for i in range(city_number):
        ratio_p = np.random.choice(PRIMARY_RATIO_LIST)
        ratio_s = np.random.choice(SECONDARY_RATIO_LIST)
        ratio_n = 100 - ratio_p -ratio_s
        j = []
        j.append((ratio_p / 100))
        j.append((ratio_s / 100))
        j.append((ratio_n / 100))
        ratio_list.append(j)

#Step2. 全都市のランク付集合ラベルリストの生成
    cities_labels_list = []
    data_number2 = DATA_NUMBER*2
    for k in range(city_number):
        l = []
        for m in range(data_number2):
            weight = ratio_list[k]
            m = np.random.choice(RANK_LIST, p=weight)
            l.append(m)
        cities_labels_list.append(l)

#Step.3 主ラベルが存在するデータで構成されるデータセットの生成
    ds1 = pd.DataFrame({
            'A1_city':cities_labels_list[0],
            'A2_city':cities_labels_list[1],
            'B1_city':cities_labels_list[2],
            'B2_city':cities_labels_list[3],
            'C1_city':cities_labels_list[4],
            'C2_city':cities_labels_list[5]
            })
    ds2 = ds1.query('A1_city == 2|A2_city == 2|B1_city == 2|B2_city == 2|¥
                    C1_city == 2|C2_city == 2')
    data_counter = len(ds2)

ds3 = ds2.sample(n = DATA_NUMBER)
ds3.reset_index()
ds3.to_csv('./dataset_labels.csv', index= False)
```

出所：筆者作成

A-2 量的データカラム追加アルゴリズムのコード

```
import numpy as np
import pandas as pd

dataset = pd.read_csv('dataset_labels.csv')
data_number = len(dataset)

#正規分布における平均値の設定(10:良い,0:普通,-10:悪い)
MEAN_LIST = [10,0,-10]
def cal():
    x = np.random.choice(MEAN_LIST)
    return x
A_mean = cal()
B_mean = cal()
C_mean = cal()

#正規分布に基づく乱数の生成(標準偏差:5)
df = pd.DataFrame({
    'A1_random':np.random.normal(A_mean,5,data_number),
    'A2_random':np.random.normal(A_mean,5,data_number),
    'B1_random':np.random.normal(B_mean,5,data_number),
    'B2_random':np.random.normal(B_mean,5,data_number),
    'C1_random':np.random.normal(C_mean,5,data_number),
    'C2_random':np.random.normal(C_mean,5,data_number)
    })

columns = df.columns.values.astype(str)
random_values = df[columns].values

cities_list = dataset.columns.values.astype(str)
labels_values = dataset[cities_list].values

num_array = np.zeros(data_number)
for i in range(data_number):
    w = np.array(labels_values[i])
    j = np.average(random_values[i], weights = w)
    j = round(j,1)
    num_array[i] = j

dataset['numerical_value'] = num_array
dataset.to_csv('./dataset_original.csv', index= False)
```

出所：筆者作成

A-3 ラベルの階層粒度変換アルゴリズムのコード

```
import numpy as np
import pandas as pd

dataset = pd.read_csv('dataset_original.csv')

A_country = ['A1_city', 'A2_city']
B_country = ['B1_city', 'B2_city']
C_country = ['C1_city', 'C2_city']

def cal_trans(x):
    array1 = dataset[x].values
    a = array1.size
    b = array1.ndim
    num = int(a / b)
    array2 = np.zeros(num, dtype ='int8')

    for i in range(num):
        a = array1[i]
        b = a.max()
        array2[i] = b
    return array2

df = pd.DataFrame({
    'A_country':cal_trans(A_country),
    'B_country':cal_trans(B_country),
    'C_country':cal_trans(C_country),
    'numerical_value':dataset['numerical_value']
    })

df.to_csv('./dataset_transformation.csv', index = False)
```

出所：筆者作成

A-4 分析工程のコード

```
import matplotlib.pyplot as plt
import pandas as pd

dataset = pd.read_csv('dataset_transformation.csv')

def cal(labels,LE,PE,PN,LN):
    LE_cal1 = dataset.query(LE)
    PE_cal1 = dataset.query(PE)
    PN_cal1 = dataset.query(PN)
    LN_cal1 = dataset.query(LN)
    ratio_LE_PE = round((len(PE_cal1))/(len(LE_cal1)),2)
    ratio_PE_PN = round((len(PN_cal1))/(len(PE_cal1)),2)
    ratio_PN_LN = round((len(LN_cal1))/(len(PN_cal1)),2)
    print(labels,
            '|LE|=' + str(len(LE_cal1)),
            '|PE|=' + str(len(PE_cal1)) + '重要度指数:' + str(ratio_LE_PE),
            '|PN|=' + str(len(PN_cal1)) + '集中度指数:' + str(ratio_PE_PN),
            '|LN|=' + str(len(LN_cal1)) + '閉鎖度指数:' + str(ratio_PN_LN)
            )
    #量的データに対する記述統計
    LE_cal2 = LE_cal1['numerical_value']
    PE_cal2 = PE_cal1['numerical_value']
    PN_cal2 = PN_cal1['numerical_value']
    LN_cal2 = LN_cal1['numerical_value']
    index_list = ['mean','50%','75%','25%','std']

    dsc_table = pd.DataFrame({
            'LE':LE_cal2.describe().loc[index_list],
            'PE':PE_cal2.describe().loc[index_list],
            'PN':PN_cal2.describe().loc[index_list],
            'LN':LN_cal2.describe().loc[index_list]
            })
    print(dsc_table)

    #箱ひげ図作成
    fig1 = plt.figure(figsize=(10,5))
    fig2 = fig1.add_subplot(1, 1, 1)
    fig2.boxplot([LE_cal2,PE_cal2,PN_cal2,LN_cal2],
                 labels = dsc_table.columns.values)
    plt.xlabel('Criteria')
    plt.ylabel('Numerical Value')
    plt.grid(True)

cal('{A,B}','A_country >= 1 | B_country >= 1','A_country == 2 | ¥
    B_country == 2','C_country <= 1','C_country  == 0')
```

出所：筆者作成

受理日 2019 年 9 月 4 日