2P1-14in1

データ分析効率化のための尺度水準判定方式

A Classification Method of Data-Scale for Improvement of Data Analysis Efficiency

平山 淳一 Junichi Hirayama
嶺 竜治 Ryuji Mine

(株)日立製作所 研究開発グループ
Hitachi, Ltd. R&D Group

ETL（Extract, Transform, Load） is absolutely essential process for bigdata analysis. In general, effectual transform operations differ by property (number, ID, text, time-clock, elapsed-time… etc) of target data. Therefore, data analysts have to pre-configure these operations before an analysis manually, and it has been a bottleneck for speeding-up of data analysis. In order to solve this problem, we proposed a classification method of “level of data scales” for reducing pre-configure cost. Out experiments on real-data illustrated over 80% classification rate of the level of data scale.

1. はじめに

データ分析による価値創出のニーズは依然として高く，大量データから素早く，業務改善に有効な分析結果を得る技術が求められている．

データ分析は ETL（Extract, Transform, Load）を基本とした工程で構成される．有効な分析結果を得るには，特にデータの変換・加工(Transform)の工程が重要であり，元データから最終的な所望のデータに至るまでに様々な変換がなされる．一例として，下記のような加工が考えられる．

- 名称の符号化（例．商品コード”4625”を 1，”4652”を 2 に）
- カラム同士の演算で新たなカラムを作成（例．単価×個数 ⇒ 売上）
- カラムを複数カラムに分割（例．7 桁の従業員 ID を，上位 3 桁のリージョン部と，下位 4 桁のシリアル部に分ける）
- 複数テーブルのマージ（例．顧客 ID をキーに POS データと顧客マスタをマージ）
- 複数行のデータの集計（例．店舗ごとの売り上げ=店舗コードが同じ行に対して，売上の合計を算出）

上記のようなデータ加工においては，対象カラムが数量なのか，ID なのか，時刻なのか，といったデータの性質の違いにより意味のある加工の種類が異なる．例えば，数量に対して平均を求める加工には意味があるが，ID に対しては意味を成さない．逆に，数量は集計のキーにはならないが，ID はなり得る，などが挙げられる．

分析時に各カラムデータの性質を適切に設定することで，意味のないデータ加工時に警告を上げる，データ集計時のキーと値を半自動で選択する，といったように，分析の有効さや素早さを高めることができると考えられる．しかし，分析対象のデータテーブルに対して事前知識がない場合，これらの事前設定が難しいといった課題があった．

我々は，この課題に鑑み，設定すべきデータの性質を，尺度水準と呼ばれる基準と紐づけ，尺度水準を自動判定することで，データ分析の効率化を行う方式に取り組んだ．以下，本稿では，尺度水準の概要，尺度水準判定方式の提案，提案方式の精度検証結果について述べる．

連絡先：(株)日立製作所 研究開発グループ システムイノベーションセンタ，junichi.hirayama.qq@hitachi.com

2. 尺度水準の種類

尺度水準とは，カラムに保存されているデータを，それらが表現する情報の性質に基づき数学・統計学的に分類する基準である．Stanley Stevens が提案した分類[Stevens 1946]がよく用いられており，本研究でもこれを用いた．尺度には低い方から順に以下の4つの水準があり，高い水準はより低い水準の性質を含む形になっている．

- 名義尺度（Nominal Scale）
  数字・文字を単なる名前として個々のデータに割り振る．2つのデータに同じ名前がついていればそれらは同じカテゴリに属する．データ間の比較は等しいか異なるかでのみ可能である．順序はなく，加減などの算術演算もできない．代表値は最頻値で表される．例としては，作業 ID，業者 ID などがある．例えば作業 ID =（00001,00002,00004,00007,…）は，作業 ID=00001 と作業 ID=00002 のデータは作業が違うことのみを表し，どちらが大きいかといった比較はできない．
- 順序尺度（Ordinal Scale）
  データに割り振られた数字・文字は順序を表す．データ間の比較は等しいか異なるかに加え，その前後・大小関係にも意味がある．一方，順序の間隔は等しくないため，加減などの算術演算には意味がない．例としては，アンケート評価やオーダー順などがある．例えば，アンケート評価 = (5, 4, 3,…)に対して，5 よりも 4 の方が良いといった比較はできる．一方，5→4 の間隔と，4→3 の間隔は均一ではなく，単純に差をとった 1 という値は意味を成さない．
- 間隔尺度（Interval Scale）
  データに割り振られた数字は順序尺度の性質を全て満たし，さらに差が等しいということは間隔が等しいということを意味する．2つのデータ間の差を比較しても意味がある．加減算にも意味があるが，尺度上のゼロ点は任意で負の値も使える．代表値は最頻値，中央値，算術平均で表される．例としては，時刻や日付などがある．例えば，日付 =(11/4, 11/6, 11/8….)に対して，11/4→11/6 の差をとった 2[日間]には定量的な意味があり，同様に 11/6→11/8 の 2[日間]との大小の比較が可能である．
- 比例尺度（Ratio Scale）
  データに割り振られた数字は間隔尺度の性質を全て満たし，さらに2つのデータの比にも，乗除算にも意味がある．尺度上のゼロ点は絶対的である．代表値は最頻値，中央

値，算術平均，幾何平均で表される．例としては，経過時間や出荷数量などがある．例えば，数量 = (2,5,10,…)に対して，2[個]と 5[個]の比をとって，2.5 倍多いといった意味づけが可能である．

図 1 に各尺度水準が持つ特徴と可能な算術操作をまとめる．

特徴 / 尺度水準	名義	順序	間隔	比例
値同士の等/不等に意味を持つ	○	○	○	○
最頻値，ユニーク数，度数分布が求められる	○	○	○	○
値同士の大小に意味を持つ		○	○	○
中央値が求められる		○	○	○
値同士の差に意味を持つ			○	○
和、差，算術平均が求められる			○	○
値同士の比に意味を持つ				○
積，商，幾何平均が求められる，原点は0となる				○

図 1 尺度水準の特徴

また，図 2 に各尺度水準のデータ例を記す．

尺度	概要	データの例
比例尺度	比率に意味あり 原点0を持つ	作業効率：(0.353, 1.246…[個/秒]) 経過時間：(14:00, 30:00…)(840, 1800…[秒]) 点数、評価指標：(54.1, 23.4, 45.2…[pt.]) 数量、大きさ：(2, 40, 210…[個])
間隔尺度	間隔に意味あり	時刻：(13:15:00…)(131500…) X座標：(0, 1, 2 …) 日時：(2013/11/23,…)(20131123, …)
順序尺度	大小関係に 意味あり	レベル、成績：(5, 4, 3, 2, 1)(A, B, C, D, E) シーケンス：(1, 2, 3, 4, 5)
名義尺度	他のデータとの 区別のみに 意味あり	ID、コード：(86100, 787813…) 名称：(日立太郎, 日立花子….) フラグ：(0, 1)(T, F)

図 2 尺度水準のデータ例

3. 尺度水準判定方式

3.1 尺度水準判定の難しさ

一般にデータの性質はデータ型によって分類されることが多い．例えば，数字(1,2,3, …)，文字列(”男”,”女”)，時間(01:20:00, 14:00:00)がある．これらは主にデータの表記を元に分類される．しかし，データの表記のみでは，正しくデータの性質を分類できないケースも多く存在する．

例えば，小売店舗の売上 POS データを考える．数字「140000」は，以下のような複数の可能性が考えられる．

- 何らかの数量(金額，売上)を示す 140,000 という値(比例尺度)
- 時刻 14:00:00 のセミコロンを省略した形式(間隔尺度)
- POS の 14 万件目を示すシーケンス番号(順序尺度)
- 商品コードとしての”140000”(名義尺度)

このように，対象データが数字の場合，データの表記的な特徴のみからは，尺度水準のようなデータの性質を特定することは難しく，その他の特徴と組み合わせる必要がある．

3.2 尺度水準判定の方針

データの文字列表記に加え，統計的性質から，尺度水準を自動判定することを試みる．統計的性質を示すものとして，度数分布を考える．図 3 は，それぞれあるカラムデータの値を横軸に，縦軸に頻度をとったヒストグラムである．図 3 左上が名義尺度である ID データ(86000 台が多い従業員 ID)，右上が順序尺度であるシーケンスデータ(作業番号，1 つの作業番号に対して複数レコードが存在する)，左下が間隔尺度である時刻データ(時刻を午前 0 時からの経過時間で表したもの)，右下が比例尺度である数量データ(作業に要した所要時間)のヒストグラムである．

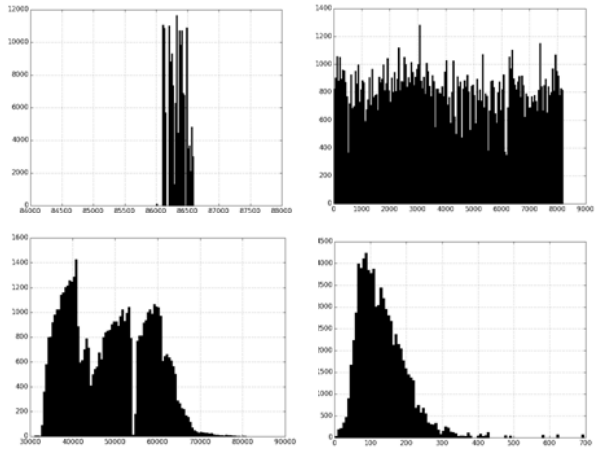

図 3 尺度水準ごとの度数分布例

これらを観察すると，一例として，以下のような分類のためのヒントが得られる．

- 名義尺度のデータでは，ある一定の範囲内(図の例では，86000 台)に不自然にデータが集中する，分布の歯抜けも多く，頻度はランダムに変化する
- 順序尺度のデータでは，一定の範囲内にデータが集中し，分布の歯抜けは少ない，頻度は一様分布に近い
- 間隔尺度のデータでは，頻度が局所的に単調増加・減少し，グラフの形状が比較的滑らかになる
- 比例尺度のデータでは，頻度グラフが何らかの中心性を持つ統計分布に似た形状になる，図の場合，対数正規分布もしくは正規分布状になる

上記はあくまで一例であるが，度数分布の特徴を利用することで，尺度水準の区別が出来る可能性を示唆している．本方式では，データの度数分布形状の傾向を示す特徴量を作成し，それら特徴量を用いて尺度水準を判定することを試みる．

3.3 尺度水準判定の特徴量

尺度水準判定のための特徴量について以下に示す．特徴量作成の観点も併せて記す．(1)(2)以外は，数値型のカラムを対象とした特徴量である．

(1) 正規表現(regex_date, regex_datetime, regex_series)

- [概要] 対象カラム内の全データが，特定の正規表現に合致した場合にフラグを立てる．本方式では日付(regex_date)と時刻(regex_datetime)，リスト表記(regex_series)を用いた．Basic Regular Expression 規格で下記のように定義した．

  regex_date：'\d{1,4}[/-]{0,1}\d{1,2}[/-]{0,1}\d{1,2}'
  regex_datetime：'\d{1,2}:{0,1}\d{2}:{0,1}\d{2}'
  regex_series：'\d{1,2}[:,\.]\D+

- [作成の観点] あるデータが日付や時刻であった場合，間隔尺度となる．また，リスト表記を持つ場合，順序尺度の可能性が高いと判断できる．

(2) ユニーク数
- [概要] 対象カラムに含まれるデータのうち，表記がユニークなものの数
- [作成の観点] 比例尺度の場合，ユニーク数が大きくなりやすい．また，ユニーク数が極端に小さい場合（例えば 5 以下），名義尺度として分析するのが妥当である．

(3) 増分の最小値（min-increment）
- [概要] データの増分値のうち，最小の値
- [観点] 最小増分幅が大きい場合，比例尺度にはなりづらい．

(4) 増分の最頻値（mode-increment）
- [概要] データの増分値のうち，最も頻度の高い値
- [作成の観点] 主に，他の特徴量の正規化に使用する．例えば時間[s]が 1 分毎に記録されている場合，増分は 60[s]になり，他の特徴量を計算する場合に 60 で割る必要がある（単位を直す）．

(5) 分散（標準偏差）/レンジ比（std/range, var/range）
- [概要] 分散(or 標準偏差)÷レンジ（最大値－最小値）で算出される値．
- [作成の観点] 名義尺度である場合，ヒストグラムが疎になりやすく，レンジに比べて分散が大きくなりやすい．従って，分散/レンジ比が大きい場合名義尺度になりやすく，小さい場合比例尺度になりやすい．

(6) 歪度（skew）
- [概要] 数値データの度数分布に対する非対称性を示す統計量．
- [作成の観点] ID 系のデータ（名義尺度）や時刻のデータ（間隔尺度）の場合，分布が一様分布になりやすく歪度が小さくなる．一方，数量データ（比例尺度）は対数正規分布に近くなるため，歪度が大きくなりやすい．

(7) 尖度（kurt）
- [概要] 数値データの度数分布に対する鋭さを示す統計量
- [作成の観点] 比例尺度である場合，値の頻度分布が単峰性の滑らかな形になりやすい．正規分布の場合尖度=3 となる．尖度 3 以上の場合，比例尺度の可能性がある．

(8) 最小値（min）
- [概要] データの最小値
- [作成の観点] データ型が整数型もしくは浮動小数点型の場合に，データ内に負値があれば比例尺度になりやすい．

(9) 増分の分散(標準偏差)（var-increment, std-increment）
- [概要] 値をユニークにした後の，全増分値（値の間隔）の分散
- [作成の観点] 名義尺度である場合，値が飛び飛びになりやすいため，増分にばらつきが出る．逆に比例尺度の場合，連続値に分布するため増分の偏差は小さくなる．

3.4 Random Forest を用いた識別

前節に示した特徴量を元に，識別器を用いた学習を行う．識別器には RandomForest[Breiman 2001]を用いた．入力は各カラムデータの特徴量リスト，正解ラベルは各カラムデータの正解尺度である．図 4 に全体フローを示す．

学習用入力テーブル

作業番号	作業者ID	処理数	時刻	商品ID	長さ	優先度
00001	84234	5	100114	203	180	1
00002	84244	9	100532	203	150	2
00003	84423	7	104323	234	150	1

特徴量テーブル

		特徴量				正解 Label
		Date	UniqNum	Skew	…	
データカラム	作業番号	0	5000	3.21	…	Ordinal
	作業者ID	1	183	0.13	…	Nominal
	処理数	0	3560	-0.02	…	Ratio
	時刻	1	4928	0.53	…	Interval
	商品ID	0	345	0.32	…	Nominal
	長さ	0	21	-2.94	…	Ratio
	優先度	0	4	5.21	…	Ordinal

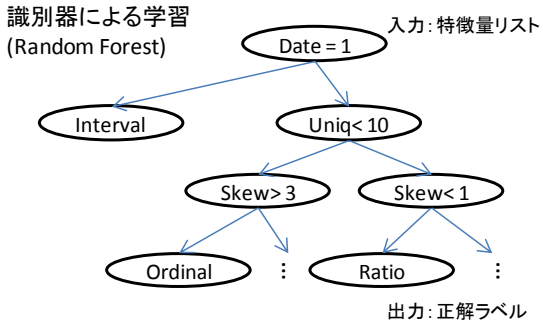

図 4 尺度水準判定フロー

4. 評価実験

4.1 データセット，評価方法

評価に用いたデータセットは，Kaggle[Kaggle]公開データ，統計言語 R の付属サンプルのほか，自社システムから収集したデータテーブルである．全 167 カラムであり，平均レコード数は 124 である．各カラムに対し，筆者の主観で正解尺度を割り振った．167 カラムのうち，8 割を学習用カラム，残り 2 割テスト用カラムとし，これを 5 回実施する交差検定法により精度の評価を行った．

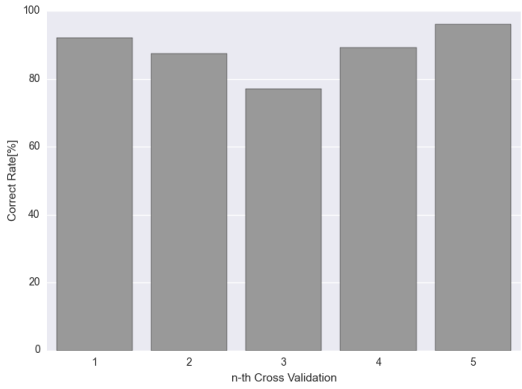

図 5 交差検定法による正解率

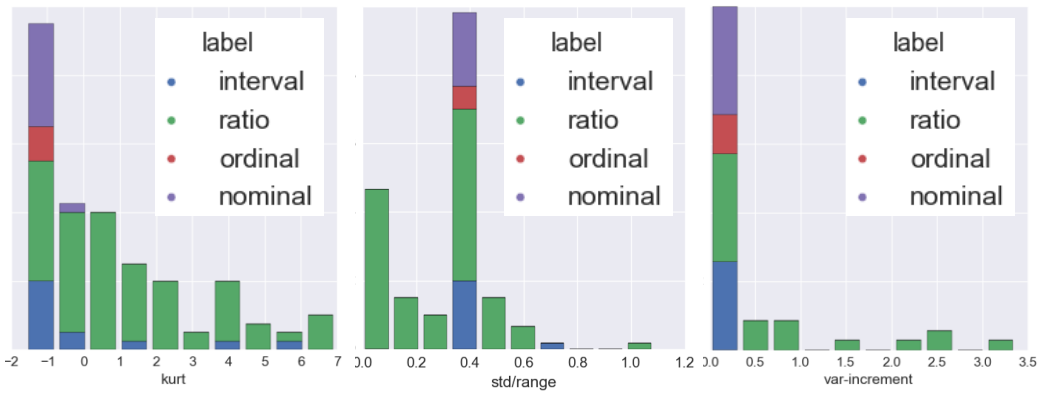

図 6 各特徴とデータ頻度の関係

4.2 実験結果と考察

交差検定法における分割パターンごとの正解率グラフを図 5 に示す. 平均で 88.46%の正解率を得た.

学習したランダムフォレストのそれぞれの決定木に対し, 識別に寄与している特徴を観察した. 主に, 整数型表記を持つデータ群を対象とした. 各特徴の考察および決定木における正解ラベルの分類数を下記に示す. 併せて, 図 6 に各特徴の値とデータ数の頻度グラフを示す.

(1) 尖度が小さい(-1.49 以下)場合に名義尺度になりやすい

- 名義尺度 ： 27/37
- 順序尺度 ： 3/20
- 間隔尺度 ： 2/19
- 比例尺度 ： 0/101

⇒ [考察] 度数分布が一様分布に近い, もしくはランダムな形状になっており, 数値として傾向を持たないデータであると考えられる. また, ユニーク数が少ないデータ(0/1 のフラグデータ等)もこれに該当しやすい.

(2) 標準偏差÷レンジ比が小さい(0.32 以下)場合に比例尺度になりやすい

- 名義尺度 ： 0/37
- 順序尺度 ： 0/20
- 間隔尺度 ： 1/19
- 比例尺度 ： 52/101

⇒ [考察] 度数分布に歯抜けが少なく, 密になっているデータと考えられる. さらに, 密なだけでなく, 度数分布が中央に偏りがあるデータと考えられる.

(3) 増分の分散が大きい(0.02 以上)場合に比例尺度、間隔尺度になりやすい

- 名義尺度 ： 5/37
- 順序尺度 ： 0/20
- 間隔尺度 ： 14/19
- 比例尺度 ： 90/101

⇒ [考察] ID(名義尺度)やシーケンス(順序尺度)データは, 値の間隔が 1 刻みになりやすい. 一方, 数量データ(比例尺度)は, 値の間隔が一律でないため, 分散が大きくなりやすい.

本実験における誤判定内訳の対応表を図 7 に示す.

順序尺度→名義尺度の誤判定が特に多かった. これらを観察したところ, 1 始まりのランキング(順序尺度)と, テーブルの主キー

		判定結果			
		名義	順序	間隔	比例
正解	名義		3		
	順序	9		3	
	間隔		2		2
	比例				

図 7 誤判定分析

である 1 始まりの ID(名義尺度)の誤判定であることがわかった. これらの両データは, ともに[1,2,3,4, …]となるデータであり, 統計的な性質は全く変わらない. これらを正しく判定するためには, 単一カラムから計算できる特徴量ではなく, 他カラムとの関連性を示す特徴量を加える必要がある. 例えばランキングの場合, スコアやポイントといった何らかの他カラムとの順位相関係数が高いと予想される. 今後の検討課題である.

4.3 今後の課題

今回の評価は, 提案方式の初期検証として, 小規模サンプルによる精度評価に留まったため, 対象サンプルデータの傾向に特異な結果になっていると考えられる. より大規模なサンプルによる評価, および更なる精度向上を継続する必要がある.

5. まとめ

本研究では, データ分析の効率化を目的とし, 入力データの尺度水準判定方式の開発に取り組んだ. 本方式により, ID, シーケンス, 数量, 時刻…etc, といったデータの性質の判定, およびデータの性質に適したデータ加工の推薦等が可能になる. 初期検証として 88.46%の判定精度を得た. 今後は, データの大規模化, および特徴量の汎化により, 更なる精度向上を進めていく.

参考文献

[Stevens 1946] Stanley Stevens: “On the theory of scales of measurement”, Science, vol.103, no.2684, pp.677-680,1946.

[Breiman 2001] Leo Breiman: “Random Forests”, Journal of Machine Learning, vol.1, issue.1, pp.5-12, 2001.

[Kaggle] https://www.kaggle.com.