表形式データにおける属性値の階層構造抽出手法の検討

辻 拓人† 引地 謙治† 宇治橋 善史† 清水 雅芳†

†株式会社富士通研究所 〒211-8588 川崎市中原区上小田中 4-1-1

E-mail: †tsuji.takuto@fujitsu.com

あらまし 近年，データ分析に基づいてさまざまなビジネス上の判断を行うことが一般的である．このとき，作成時にはデータ分析に利用することを想定していなかった過去の資産を活用することが求められている．しかし，そのようなデータは表計算ソフトで作成されたスプレッドシートが多く，それらは作成者，部署等によってデータ構造に微妙な差異が存在する．それらを分析するために，個々のシートに合わせたデータ抽出プログラムを作成するのでは膨大なコストがかかってしまう．そこで，われわれは機械学習を活用した汎用性の高いデータ抽出技術の開発に取り組んでいる．シートからのデータ抽出に関して，Chen らはセルペアの特徴量に基づいてシート内ヘッダ領域の属性値の親子関係を自動的に抽出する方式を提案している．しかし，この方式ではヘッダ領域が一列のシートしか想定していない．そこで本稿では複数列のヘッダ領域に対応できるように特徴量を拡張する．さらに，テーブル上のセルの位置やセルペアの位置関係などに着目して特徴量を追加したときの親子関係の推定性能を評価した．

キーワード 表形式データ，スプレッドシート，属性値，階層構造，機械学習，自動化，前処理，オープンデータ

1. はじめに

現在，大量のデータを分析して企業・団体等の組織上のさまざまな判断に活用することが求められている[1]．特に組織に過去から蓄積されたデータは長年の活動によって得られた多くの知見が埋もれており，それらを活用することは組織活動に有益である可能性が高い．しかし，そのようなデータは往々にして作成時点の用途以外に活用することを想定しておらず，必ずしもデータ分析に適した形式でないことが課題となっている．そのようなデータには表計算ソフトで作成されたスプレッドシートが多く含まれる．スプレッドシートはデータ管理や分析の非専門家でもテーブルデータを容易に作成できる利点からさまざまな組織で広く作成されている．しかし，それらはデータ作成者や部署といった限定された範囲の利用しか想定されてないことが多い.そのため,データ構造が不明瞭であったり，類似データでも構造に細かい差異が存在したりする．このようなスプレッドシートを分析に活用するにはデータ構造の抽出が必要になる．しかし，個々のスプレッドシートに合わせた抽出プログラムを作成したり，人手でデータを整形したりする方法では膨大なコストがかかってしまう．さらに作成者の異動等により，データ構造の確認が困難になることも深刻な問題である．データ活用においては前準備が全体の 50 から 90%を占めると言われているが[3][4][5]，このようなデータ構造の抽出が必要なことがその一因として考えられる．

そこでわれわれはスプレッドシート内のテーブルのデータ構造を抽出する汎用性の高い方式の実現に向けた研究開発に取り組んでいる．本稿では抽出プロセスの中で特にテーブルのヘッダ領域内における属性値の階層構造推定について報告する．階層構造とはヘッダ内の属性値間における集約関係もしくは上位概念・下位概念といった構造である．例えば図 1 の 15 行目に着目すると,左側ヘッダの「農業」は「農業，林業」の一部であり,さらにそれは「全産業」の一部という階層化された集約関係が存在する．この階層構造を同図左部のように表全体に対して推定するのがここでの問題である．

Chen [2] は階層構造推定をセルの組合せ（セルペア）から生成した特徴量に基づいてそれらが親子関係であるか否かの二値分類問題に帰着させて機械学習を適用する方式を提案している．しかし，この方式ではヘッダ領域が一列のケースしか想定されておらず，複数列のケースにおける有用性は分かっていない．なお，一般的なテーブルでは左側と上側にヘッダ領域が存在するが,本稿では特に左側のヘッダに限定して議論する．また,この方式による推定性能は F1 スコアで 0.8 程度であり，作業量の効率化の観点で不十分であると考えている．われわれは，Chen の方式で提案されている特徴量をベースラインとして，複数列のヘッダを持つテーブルに対応させるなど特徴量を拡張した．その上でデータカタログサイト data.go.jp [6]，政府統計の総合

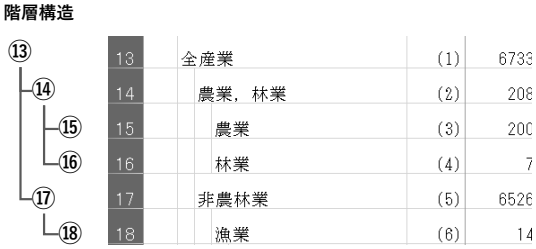

図 1 ヘッダ階層構造の例

出典：「労働力調査 / 基本集計 全都道府県 結果原表 全国」（総務省統計局）

窓口 e-Stat [7], DECO [8] から取得した複数列のヘッダを持つスプレッドシートを含むデータに対して Chen の方式を評価する．さらに Chen の方式をベースラインとしていくつかの観点で特徴量を拡張することで効果的な特徴量について検討した．本稿の構成は以下の通りである．2 章ではスプレッドシートのデータ構造抽出に関する従来技術を述べる．それらの中で特に Chen のセルペアに基づくの階層構造推定方式を詳述する．3 章では Chen の方式に対する特徴量の拡張方法ついて述べる．4 章でベースラインおよび特徴量を拡張したときの推定性能の評価方法および評価結果を説明する.最後に 5 章でまとめと今後の課題について述べる．

2. 従来技術

スプレッドシートからのデータ構造抽出は主にスプレッドシート内のテーブル領域の同定およびテーブル内のレイアウト（データ領域，ヘッダ領域など）の推定[2][10][11][12][13][14], テーブル内の値(数値データなど）とヘッダ領域の属性値の対応付け[2][12][15][16]に分類できる[9]．本稿で対象とする属性値の階層構造推定は値と属性値の対応付けにおける主要な処理である．また具体的な方式の観点ではヒューリスティクスアルゴリズム[13][15]，機械学習[2][10][11][12]に分けられる．さらに機械学習方式には Active Learning により人の判断を活用する方式[14] も存在する．これらの中で，ヘッダ領域の属性値の階層構造に着目している研究には Chen [2]，Sigarov [15], Goto [16] がある．Sigarov [15] は階層構造推定プログラムの開発工数削減を目的としたルールベースのドメイン特化言語を提案している．しかし，このアプローチではスプレッドシートの作成者や所属部署等による細かな違いに対応するためにルール数が膨大になる．さらにルール数の増加に従ってルール間の整合性維持が課題となる. Goto [16] は主に統計データを対象としてテーブル内の数値データ間の関係性に基づいて属性値の階層構造を推定する方式を提案している. 例えば，「全国」という属性値を持つ行がある場合，その下に続く「北海道」,「青森」といった都道府県の行の値の合計値が「全国」行の値になるといった関係性から階層構造を推定する．しかし，このような特徴を持つスプレッドシートに適用範囲が限定される問題がある．われわれは機械学習を用いることで作成者や部署等による細かな形式の違いに適応するコストが低く，なおかつさまざまな分野のスプレッドシートに対応可能な Chen [2] の研究に注目した．この研究ではセルペアの特徴量に基づく機械学習を用いた階層構造推定方式 Hierarchy Extractor を提案している．この方式は階層構造推定をヘッダ領域内のセルペアが親子関係であるかを分類する二値分類問題に帰着させている．これは以下のように定式化される：

定義（Hierarchy Extractor） $A = \{a_1, \ldots, a_N, root\}$ をテーブルヘッダ領域内のセルの集合とする．このとき，ある $a_i, a_j \in A$ に対してセルペア (a_i, a_j) が ParentChild であるとはa_iがa_jの親であることを表す．階層構造推定とは全てのセルペアの中で ParentChild であるセルペアを推定することである．なお，root は仮想的な最上位セルである．

具体例として図 1 において (「農業，林業」セル,「農業」セル) は “ParentChild である”，(「農業，林業」セル,「非農林業」セル) は “ParentChild でない”，(root,「全産業」セル) は “ParentChild である” となる．Hierarchy Extractor ではまず学習用のスプレッドシートのセルペアから特徴量を算出し，特徴量と ParentChild であるか否かを表すラベルを分類器に学習させる．推論フェーズでは未知のスプレッドシートのセルペアの特徴量を分類器に入力してそのペアが ParentChild であるか判別することで階層構造を推定する．この方法で推定した結果，子セルに対して複数のセルが親セルと判定されることで階層構造が木にならないケースが発生する．そのような場合，親である確率が最も高いセルのみを親セルにすることで木構造を強制する方法も提案されている．

3. 提案方式

Chen [2]の方式ではセルペアの特徴量はヘッダ領域が一列のテーブルを想定しており，図 2 のように複数列のヘッダを持つケースに対応できない問題がある．また，事前調査により従来方式の特徴量では今回の分析対象データの特徴を十分に表現できない可能性があることから特徴量の拡張を行った．本章ではまず特徴量の分類を整理した上で特徴量の拡張について述べる．

3.1. 特徴量の分類

Chen [2]はセルペアの特徴量をセルペアの中のセル単体から決まる Unary 特徴量と 2 つのセルから決まる Binary 特徴量に分類している．本稿ではさらに Binary

		製造業	0.19		0.16
		農林漁業			
		鉱業，採石業，砂利採取業，建設業	0.54		-0.12
	非	電気･ガス･熱供給･水道業			
民		運輸業，郵便業	0.23	-0.16	-0.28
	製	情報通信業	0.17		
間		卸売業，小売業	-0.11	-0.19	-0.16
	造	金融業，保険業	-0.46		
等		不動産業	-1.16	-0.32	-0.37
	業	サービス業	-2.39		
		その他			
		小計	-3.18	-0.49	-1.04

図 2 複数列のヘッダの例

出典：「建設工事統計調査 / 建設工事受注動態統計調査 / 大手５０社」（国土交通省）

特徴量の中で 2 つのセルの中間領域に関する特徴量を Middle 特徴量に分類する（それぞれ Appendix A の表 3, 表 4, 表 5）．さらに，各分類の中では，「合計という文字を含む」といったセルのテキストに関する意味的特徴量，「テキストが bold である」といったフォントや背景色等に関する装飾的特徴量および，「2 つのセルが隣接している」といったスプレッドシート上の配置に関する構造的特徴の観点でも分類できる．

3.2. 特徴量の拡張

Chen [2] における特徴量に対して以下の拡張を行った：

拡張 1 セル座標の追加（表 3 の U13, 14）

拡張 2 装飾的特徴の追加（表 3 の U11, 12，表 4 の B6, B7）．B6, B7 はスタイルの定義に U11, U12 の定義を加えている．

拡張 3 複数列のヘッダに対応した Middle 特徴量の領域定義．表 5 の M1-M6 の定義は Chen [2]と同等であるが，ペア間の中間領域の定義が異なる．また，M7-9 は拡張 2 にあわせて追加している

以下，それぞれの拡張の意図を述べる．拡張 1 はさまざまな構造的特徴の表現が目的である．すなわち，「2 つのセルが隣接している」という位置関係の特徴を考えると，これは「左右に隣接」,「上下に隣接」に分解できるが，このようなバリエーションを試行錯誤して追加するのは汎用性の観点で望ましくない．それに対して座標情報を与えれば位置関係に関する特徴を分類器が自動的に決定することを期待できる．

拡張 2 は対象データの事前調査に基づいて，従来方式で不足している装飾特徴量を追加したものである．

拡張 3 は従来方式が複数列のヘッダを持つテーブルを想定していない問題の解決を目的としている．Middle 特徴量の算出において，ヘッダが一列であれば 2 つのセルの中間領域を一意に定義できるが，複数列の場合はいくつかの定義が考えられる．そこで効果的な中間領域の調査を目的として以下の通りに中間領域を定義した：

中間領域 1 2 つのセルに挟まれた全てのセル

中間領域 2 2 つのセルを両端とした L 字領域に存在するセル

以下にそれぞれの定義の詳細と具体例を示す．

定義（中間領域 1） セルa_iの行インデックス，列インデックスをそれぞれr_i, c_iとする．このときセルペアa_1, a_2 $(r_1 \le r_2)$ に対する中間セルの集合M_{all}を

$$M_{all} = \{a_i \mid a_i \in A \ \wedge m_{a_i} = 1\}$$

$$m_{a_i} = \begin{cases} 1 \ \text{if} \ r_i = r_1 \ \wedge c_i > c_1 \\ 1 \ \text{if} \ r_1 < r_i < r_2 \\ 1 \ \text{if} \ r_i = r_2 \ \wedge c_i < c_2 \\ 0 \ \text{otherwise} \end{cases}$$

とする．例えば，図 3 の塗りつぶされたセルがセル 1 とセル 2 の中間セルである．

定義（中間領域 2） セルペアa_1, a_2 $(r_1 \le r_2, c_1 \le c_2)$ に対する L 字型中間セルの集合M_{LL}を

$$M_{LL} = \{a_i \mid a_i \in A \ \wedge l_{a_i} = 1\}$$

$$l_{a_i} = \begin{cases} 1 \ \text{if} \ r_1 < \ r_i < r_2 \ \wedge c_i = c_1 \\ 1 \ \text{if} \ r_I = r_2 \ \wedge \ c_1 \le c_i < c_2 \\ 0 \ \text{otherwise} \end{cases}$$

とする．具体例を図 4 に示す．この L 字型の領域は人が親子関係を確認する際，まず上位の親のセルを見て，その後下の行を辿りつつ右側のセルを見て 2 つのセルが親子関係か判断することが多かったことから，親子関係の判断に関わる情報が含まれるという観点で定義した．

4. 評価

複数列のヘッダを持つテーブルに対する推定性能および，前章で述べた特徴量の拡張効果の確認を目的としてオープンデータを用いた評価を行った．本章では評価方法を説明し，その後に分類器の選択を目的とした事前実験について述べて，最後に特徴量の違いに対するセルペアの ParentChild 推定性能の評価を示す．

4.1. 評価方法

評価の詳細として，利用したデータ，評価手順を述べる．

データ e-stat, data.go.jp から複数のヘッダ列を持つテーブルを含む 73 のスプレッドシート（以下，単にシートと呼ぶ）を選択した．1 つのシートには 1 つのテーブルが含まれている．これらのデータは主に統計データであり，留意すべき特徴として一つの表の中で類似する階層構造が繰り返し現れることが挙げられる．例えば図 1 の (「農業，林業」セル,「農業」セル) と，(「非農林業」セル,「漁業」セル) のようなケースであり，これらのセルペアはセルの絶対的な位置とテキス

図 3 中間領域 1 の具体例

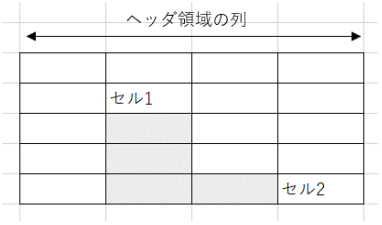

図 4 中間領域 2 の具体例

トを除けば基本的に特徴量が同じになる.

正解データの作成 表の各行に対応付く順序付けられた属性値を付与しそれをセルペアに対する ParentChild ラベルに変換することで作成した．例えば図 1 の 15 行目に対する属性値は階層構造のルート側から「全産業」,「農業，林業」,「農業」であるから，“ParentChild である” セルペアとして (「全産業」,「農業，林業」), (「農業，林業」,「農業」) の二つができる．これ以外のセルペアは全て“ParentChild でない” とする.なお，シート内のテーブル位置およびヘッダ領域は手動で決定した.

評価手順 セルペアから作成した特徴量と ParentChild ラベルのセット（以下，データセットと呼ぶ）を学習データと評価データに分割して分類器の学習と推定性能の評価を行う. 評価値としては Precision（適合率），Recall（再現率），F1 スコアを用いた．また，データセットの分割方法は評価対象ごとに異なるので次節以降で説明する.

4.2. 分類器の評価

はじめに分類器の選択を目的とした評価を説明する．分類器の候補としてランダムフォレスト，SVM, ロジスティック回帰を用いた．それぞれ機械学習ライブラリ scikit-learn [16] の RandomForestClassifier, SVC, LogisticRegression クラスを初期設定で利用している．この評価ではデータセットを以下のように分割して評価した：

データセット分割方法 10 シートのデータセットを統合して 75%を訓練データ, 25%を検証データに分ける.

全シートのデータセットを統合するのは，シートごとの特徴の違いが推定性能に与える影響を緩和するためである．分類器ごとの評価結果を表 1 に示す．特徴量は次節で説明する特徴量セット F を用いた.これよりいずれの評価値でもランダムフォレストが最もよい性能であることが分かった.また分類器の動作の分析が容易な利点もあることから，以降の評価ではランダムフォレストのみを利用する．なお，評価結果では F1 スコアで 0.8 以上となっているが，この評価では訓練データと評価データに同じ特徴を持つセルペアが含まれている影響が大きい.そのため未知データに対する評価として妥当でないことを指摘しておく.

4.3. 特徴量の拡張の評価

3.2 節で述べた特徴量の違いに対する推定性能を評価する．この評価ではデータセットを以下のように分割して評価した：

データセットの分割方法 1 つのシートのデータセットを訓練データと評価データに均等に分割して，シートごとに評価値（F1 スコア，Precision, Recall）を求める.

表 1 分類器の性能評価

分類器	F1	Precision	Recall
ランダム フォレスト	0.866	0.918	0.820
SVM	0.575	0.819	0.443
ロジスティック 回帰	0.535	0.875	0.385

表 2 評価 2 の測定結果

特徴量 セット	F1	Precision	Recall
A	0.410	0.638	0.369
B	0.705	0.848	0.651
C	0.519	0.702	0.489
D	0.506	0.704	0.465
E	0.518	0.707	0.482
F	0.757	0.860	0.714

このように分割した理由は特徴量の違いが分類性能に与える影響にフォーカスして分析するためである．すなわち，同一シートのテーブルには類似する階層構造が繰り返し含まれる可能性が高いことから，同じシートの一部のデータセットで分類器を訓練してその分類器で分類性能を評価した結果，十分な性能が出ない場合は特徴量がセルペア本来の特徴を正しく表現できてない可能性が高いと推測できる．また，後で述べるがシートごとに分類性能が大きくばらつくため，シートの違いが分類性能に与える影響を分析することも目的としている．今回評価した特徴量セットの詳細は以下の通りである.

- **特徴量セット A（ベースライン）** Unary 特徴量と Binary 特徴量(座標に関する特徴 U13,U14 を除く)
- **特徴量セット B** ベースラインに拡張 1(セル座標) を追加
- **特徴量セット C** ベースラインに拡張 2(装飾的特徴) を追加
- **特徴量セット D** ベースラインに拡張 3(中間領域 1) を追加
- **特徴量セット E** ベースラインに拡張 3(中間領域 2) を追加
- **特徴量セット F** ベースラインに拡張 1 拡張 2, 拡張 3（中間領域 1）を追加

特徴量セット B から E は特徴量の拡張 1,2,3 それぞれのベースライン（セット A）と比較した効果を分析することを目的としている．F については全ての拡張を含めた総合的な評価としている.

各特徴量セットに対する評価結果を表 2 に示す. F1, Precision, Recall の値は 73 シートそれぞれに対する評価結果の平均値を示している．この結果より，ベースラインに拡張 1-3 のいずれかを加えた特徴量セット B-E は，どれもベースラインより推定性能が向上することが分かった．特に拡張 1（セル座標）を追加したセット B がそれらの中でもっとも性能がよい．また中間領域に関する拡張についてはセット D, E いずれもベースラインよりは性能が向上するが中間領域 1, 2 の間で大きな差はみられなかった. また，拡張 1,2,3 全てを追加したセット F は座標のみを追加したセット B よりも性能が向上した.

4.4. 個別シートの評価

図 5，図 6 に特徴量セット A, B のケースの F1 スコアの度数分布を示す．同図より，特徴量セット B はベースラインより平均としては性能が向上しているが，個別のシートの評価値は大きくばらつくことが分かる. そこで個別のシートに対して特徴量を追加したこと影の響を調査した.

ベースラインでは親子関係を正しく判定できなかったが，特徴量の拡張 1（セル座標追加）によって改善された例を図 7 を用いて説明する. 図 7(a) のセルペアはベースラインでは誤って “ParentChild でない” と判定した (false negative)ケースである．また同図(c)はベースラインでは誤って “ParentChild である” と判定した (false positive) ケースである．いずれも拡張 1 により正しく判定された．これらのセルペアは装飾情報が少なく座標情報により位置関係が親子関係の検出に効果的に寄与したと考えられる．また，同図(b)はいずれの特徴量セットでも誤って “ParentChild でない” と判定した (false negative) ケースである. このように空白や罫線によって表現された結合領域が存在する場合は事前に結合領域を検出してそれを一つのセルとみなすことで推定性能が改善される可能性がある.

また，ベースラインに対して拡張 3（中間領域）によって推定性能が大きく改善される個別シートは存在しなかった．これは中間領域に対して検査する内容に問題があった可能性がある．例えば “中間領域にテキストを含むセルがある” という特徴量は図 8(a)のセルペアでは値が True になるが，同図(b)の場合は False になるというケースである．このように中間領域の検査内容について今後の検討課題である.

5. おわりに

本稿では組織に埋もれているデータを分析に活用することを目的として，スプレッドシート内のテーブルヘッダにおける属性値の階層構造の自動抽出方法を検討した．従来技術ではヘッダが一列のテーブルを想定していなかった問題に対して，複数列に対応させたり，セルペアの中間領域を考慮したりして特徴量を拡張した．しかしながら，現状では十分な推定性能を得られてない．それに対してテーブル内における結合領域の抽出，自然言語処理を活用したテーブルの意味的特徴の抽出などが考えられる. また, Active Learning 等により人の判断結果を取り入れるなどシステム全体として推定性能の向上に取り組む予定である.

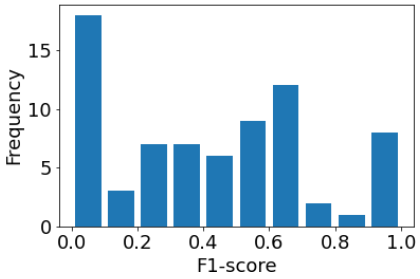

図 5 特徴量セット A の F1 スコアの度数分布

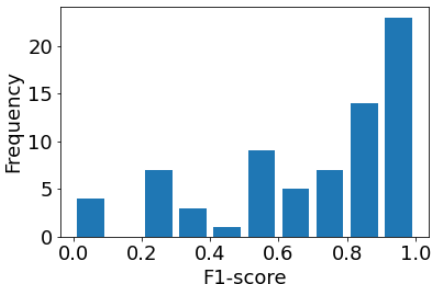

図 6 特徴量セット B の F1 スコアの度数分布

謝辞

データセットの作成に協力いただいた富士通九州ネットワークテクノロジーズ株式会社の萩原克守氏，岩瀬なみ氏，村川美樹氏に感謝する.

参 考 文 献

[1] DAMA International, データマネジメント知識体系ガイド 第二版. 日経 BP, 2018.

[2] Z. Chen and M. Cafarella, “Automatic web spreadsheet data extraction,” in Proceedings of the 3rd International Workshop on Semantic Search Over the Web, Riva del Garda, Italy, Aug. 2013.

[3] S. Lohr, “For Big-Data Scientists, ‘Janitor Work’ Is Key Hurdle to Insights,” The New York Times. https://www.nytimes.com/2014/08/18/technology/for-big-data-scientists-hurdle-to-insights-is-janitor-work.html (accessed Dec. 11, 2020).

[4] DSAIL, “Data Civilizer: A Tool to Find, Ingest, Clean, and Integrate Diverse Data Sets.” http://dsail.csail.mit.edu/index.php/data-civilizer/ (accessed Dec. 11, 2020).

[5] B. Hayes, “How do Data Professionals Spend their Time on Data Science Projects?,” BUSINESS BROADWAY. https://businessoverbroadway.com/2019/02/19/how-do-data-professionals-spend-their-time-on-data-science-projects/ (accessed Dec. 11, 2020).

[6] https://www.data.go.jp/

[7] https://www.e-stat.go.jp/

[8] E. Koci, M. Thiele, J. Rehak, O. Romero, and W. Lehner, “DECO: A dataset of annotated spreadsheets for layout and table recognition,” in Proceedings of the International Conference on Document Analysis and Recognition, ICDAR, Sep. 2019, pp. 1280–1285.

[9] R. Rastan, H. Y. Paik, and J. Shepherd, “TEXUS: A task-based approach for table extraction and understanding,” in *DocEng 2015 - Proceedings of the 2015 ACM Symposium on Document Engineering*, Sep. 2015, pp. 25–34.

[10] C. Christodoulakis, E. B. Munson, and M. Gabel, “Pytheas pattern-based table discovery in CSV files,” Proceedings of the VLDB Endowment, 2020.

[11] H. Dong, S. Liu, S. Han, Z. Fu, and D. Zhang, “Tablesense: Spreadsheet table detection with convolutional neural networks,” in *Proceedings of the AAAI Conference on Artificial Intelligence*, 2019, vol. 33, pp. 69–76.

[12] H. Dong, S. Liu, Z. Fu, S. Han, and D. Zhang, “Semantic Structure Extraction for Spreadsheet Tables with a Multi-task Learning Architecture,”, in Workshop on Document Intelligence (DI 2019) at NeurIPS 2019, 2019,

[13] E. Koci, M. Thiele, W. Lehner, and O. Romero, “Table recognition in spreadsheets via a graph representation,” in *Proceedings 13th IAPR International Workshop on Document Analysis Systems, DAS 2018*, 2018, pp. 139–144.

[14] J. Gonsior, J. Rehak, M. Thiele, E. Koci, M. Günther, and W. Lehner, “Active Learning for Spreadsheet Cell Classification,” in Proceedings of the EDBT/ICDT 2020 Joint Conference, 2020.

[15] A. O. Shigarov, V. V. Paramonov, P. V. Belykh, and A. I. Bondarev, “Rule-Based Canonicalization of Arbitrary Tables in Spreadsheets,” in Information and Software Technologies, 2016, pp. 78–91.

[16] K. Goto, Y. Ohta, H. Inakoshi, and N. Yugami, “Extraction Algorithms for Hierarchical Header Structures from Spreadsheets,” in *EDBT/ICDT Workshops*, 2016, pp. 179–188.

[17] https://scikit-learn.org/

Appendix A 特徴量の定義

セルペアから作成する特徴量を示す．表 3 の Unary 特徴量はセル単体に対する特徴，表 4 の Binary 特徴量はセルペアに対する特徴，表 5 の Middle 特徴量は中間セル全体に対する特徴である．Middle 特徴量は中間セル集合の定義とは独立に定義している．

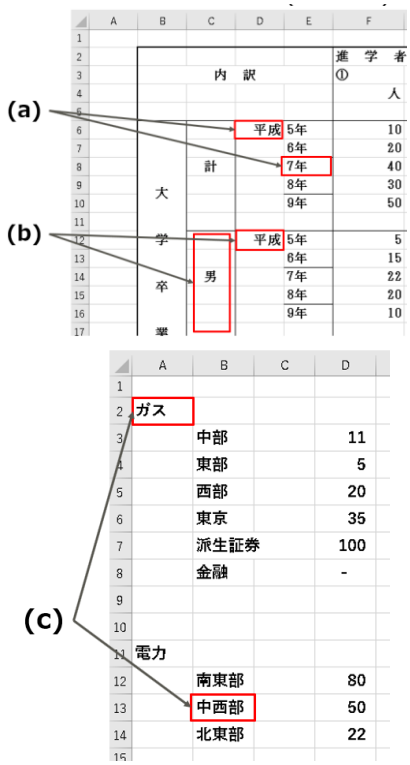

図 7 (a), (c) は拡張 1 で正しく親子関係を判定できたペア，(b) いずれの方式でも正しく判定できなかったペアの例（いずれも実際のシートの数値，文字列のみを変更して作成）

神奈川	川崎市	A工場
	横浜市	Bビル
東京	港区	Xシティ
	港区	Yセンター
	大田区	Z工場

(a)

神奈川	川崎市	A工場
	横浜市	Bビル
東京	港区	Xシティ
	港区	Yセンター
	大田区	Z工場

(b)

図 8 (a) “ParentChild である”ケース (b) “ParendChild でない”ケースの L 字中間領域（太字がセルペア）

表 3 Unary 特徴量

U1	テキストにアンダーラインがある
U2	テキストが「合計」のような単語を持つ
U3	テキストがコロンを含む
U4	テキストが bold
U5	テキストが中央寄せ
U6	テキストが italic
U7	テキストが数字
U8	テキストのアルファベットが全て大文字
U9	セルがヘッダ領域の最初の列
U10	セルがヘッダ領域の最後の列
U11	白色の背景を持つ†
U12	境界線を持つ†
U13	セルの列インデックス‡
U14	セルの行インデックス‡

表 4 Binary 特徴量

B1	ペアが隣接している
B2	ペアは同じインデント数を持つ
B3	子セルのフォントサイズが親セルよりも小さい
B4	子セルは親セルよりもインデント数が少ない
B5	子セルの行インデックスが親セルのインデックスよりも大きい
B6	子セルが 1 列目の要素のスタイルが同じ†
B7	ペアのスタイルが同じでなる†
B8	親セルがルートセルである

表 5 Middle 特徴量

M1	ペア間に空白のセルが存在する
M2	ペア間にインデントを持つセルが存在する
M3	ペア間にペアより大きなインデントを持つセルが存在する
M4	ペア間にペアより小さいインデントを持つセルが存在する
M5	ペア間にペアとスタイルが異なるセルが存在する
M6	ペア間に「合計」のような単語を持つセルが存在する
M7	ペア間に背景が白色でないセルが存在している†
M8	ペア間のセルのスタイルが全て同じ†
M9	ペア間のセルに境界線を持つセルが存在する†

注：表 3 から表 5 で†,‡が付与された項目はわれわれが独自に追加した項目を示している