The 32nd Annual Conference of the Japanese Society for Artificial Intelligence, 2018

マルチラベル分類と知識ベースの埋め込みを用いたテーブル中の列の概念決定

Determination of Column Concepts in Tables using Multi-label Classification and Knowledge Base Embeddings

竹岡 邦紘 *1 Kunihiro Takeoka　小山田 昌史 *2 Masafumi Oyamada　中台 慎二 *2 Shinji Nakadai　岡留 剛 *1 Takeshi Okadome

*1関西学院大学大学院理工学研究科 Graduate School of Science and Technology, Kwansei Gakuin University

*2NEC データサイエンス研究所 NEC Data Science Research Laboratories

Because relational data represented in tables are not connected with a knowledge base, efficient usage of tables such as semantic search in the web requires the determination of column concepts, that is, annotating a column with an entity in the knowledge base. This paper proposes a new method for annotating columns with entities in the knowledge base using multi-label classification independently. The method also annotates columns in a table with entities using knowledge base embeddings. It uses multi-label classifiers for named-entity(NE) and literal columns. Assuming Markov random field, in which entities corresponding to columns are represented by latent variables, the method maximizes the joint distribution of the latent and observed variables to estimate the optimal entities. Experiments using 183 annotated tables in UCI Machine Learning Repository with WordNet show the superiority of the method.

1. はじめに

構造化された関係データの表現であるテーブルは，一般に体系化された知識ベースとは接続されておらず，テーブルの意味的な索引付けや検索といったテーブルの活用にはテーブルの概念決定が必要になる [Venetis 11, Deng 13, Mulwad 13]．一般に，テーブルの概念決定は，以下の 3 つの対応付けからなる [Zhang 17]．

- テーブル中の各要素に対する，知識ベースにおける要素の概念を表現するノードへの対応づけ．
- テーブル中の各列に対して，1 つの列中の各要素が同一の概念の「実現値」であると仮定し，知識ベースにおけるその概念を表現するノードへの対応づけ．
- テーブル中の複数の列に対して，それらの概念的なつながりの知識ベースに含まれる関係との対応づけ．

本稿では，知識ベースのノード（1 つのエントリ）のことを**エンティティ**と呼び，列に対応する概念の決定，すなわち列に対応するエンティティ推定を扱う．

本研究では，Zhang の手法 [Zhang 17] に従い，テーブル中の列を NE 列とリテラル列の 2 種類に分ける．**NE 列**は列内の各要素がエンティティと対応づけられる列と定義し，**リテラル列**は列内の各要素がエンティティと対応づけられない列と定義する．例えば，NE 列は国名や地域名を要素とする列であり，リテラル列は国や地域の GDP を要素とする列である．つまり，リテラル列は数値とその単位，記号などを要素とする列である．定義よりリテラル列については各要素とエンティティの対応づけを行わない．

本研究では，(1) マルチラベル分類に基づいて，列ごとに列に対応するエンティティを推定する手法と，(2) リテラル列を含む複数の列間の相互依存関係を利用した，複数列に対して同時に列に対応するエンティティを推定する手法を提案する．各列に対応するエンティティは複数存在することを仮定し，列に対応するエンティティ推定をマルチラベル分類によって定式化する．また，列間の相互依存関係をマルコフ確率場によって表現し，列に対応するエンティティを表す潜在変数，タイトルを表す潜在変数，テーブルの要素を表す観測変数の同時確率を最大化することで，テーブル中の複数の列に対応するエンティティを同時に推定する．

リテラル列に対するエンティティ推定の先行研究として，Neumaier らの手法 [Neumaier 16] や Pham らの手法 [Pham 16] がある．Neumaier らは数値のみで構成される列に対して，背景知識のグラフを用意することでエンティティ推定する手法を提案している．一方で，Pham らはリテラル列を含む一般の列に対して，類似度に基づいた 2 クラス分類器を用いてエンティティ推定する手法を提案している．リテラル列は数値のみで構成されると限定されないことから，本研究では，リテラル列に対するエンティティ推定の既存手法として Pham らの手法を比較対象とする．

また，Limaye らは，マルコフ確率場によってテーブル中の要素間や列間の関係をモデル化する方法を提案している [Limaye 10]．それは，類似度を用いてポテンシャル関数を表現する自由度の高いモデルである．しかし，NE 列にのみに対応しており，リテラル列を推定するためには拡張が必要である．本研究では，Limaye らの手法をリテラル列に対して拡張したモデルをベースラインとし，その上で，列に対応するエンティティ推定に対してマルチラベル分類を組み込んだ新たなモデルを提案する．提案手法では先行研究同様マルコフ確率場を用いる．

先行研究に対する本研究の優位性は以下である．(1) NE 列とリテラル列のそれぞれの列に対応するエンティティ推定をマルチラベル分類として定式化することで，各列の特徴ベクトルと複数のエンティティとの対応づけを学習データから学習でき，さらに列に複数のエンティティが対応するような曖昧性を考慮することができるため，他の列との依存関係によってその曖昧性を解消するモデルへと接続しやすい．(2) 列間の関係を全結合関係とし，新たにタイトルという潜在変数を用いたテーブルのモデル化をすることで，列間を独立にみることでは得られない列間依存関係を列のエンティティ推定に利用できる．

連絡先: 竹岡邦紘，関西学院大学大学院理工学研究科，kunihiro.takeoka.93@gmail.com

(3) 列間の相互依存関係を含んだモデルの同時確率を最適化することで，テーブル内の複数列に対応するエンティティを同時に求めることができ，さらに列ごとに推定するときよりも性能向上する．

本研究では，提案手法を，UCI Machine Learning Repository にある 183 の CSV データと，知識ベースとして WordNet を用いて検証する．CSV データを 1 つのテーブルとみなして人手によるアノテーションを行いそれを真値とした．

2. 関連研究

2.1 類似度に基づく列に対応するエンティティ推定手法

Pham らは，テーブルの列に対応するエンティティ推定のために，類似度に基づく分類手法を提案している [Pham 16]．Pham らの手法は，各列ごとに独立にエンティティ推定を行う．1 つの列を要素の集合として表現し，集合間に文字列的な類似度と数値的な類似度を定義する．例えば，文字列的な類似度に要素の 3-gram の jaccard 係数を用い，数値的な類似度にはコルモゴロフ-スミノフ検定の P 値を用いる．このように定義された複数の類似度の値で構成されるベクトルを入力とし，それらが同じエンティティを指すかどうかを出力する 2 クラス分類器を学習する．この手法では，推論時に，新たに入力された列と学習データ内の列との類似度ベクトルを求める必要がある．従って，学習データが多ければ推定の性能は向上するが，推論にかかる時間は学習データ数に比例する．

2.2 マルコフ確率場によるテーブルの変数間表現

マルコフ確率場を用いてテーブルの変数間の関係を表現したテーブルの概念決定の手法として，Limaye らの手法がある [Limaye 10]．Limaye らは，テーブルを複数の NE 列からなるものとして，テーブルの概念決定を行う手法を提案している．要素と列名，列間の関係についてそれぞれ潜在変数を定義し，マルコフ確率場によってテーブル中の変数間の関係を表現する．各変数間のポテンシャル関数には，テーブルの要素の値である文字列とエンティティの類似度と，エンティティ間の類似度，エンティティの組と関係の間の類似度を用いる．その上で，ポテンシャル関数の積で表現される同時確率の最大化によってテーブルの概念決定を行う．また，同時確率を最大とする潜在変数の値の組を推定することは NP 困難であるため，近似推論の手法も提案している．この手法は，リテラル列を無視することを前提としており，すべてのテーブルに用いることはできない．

3. マルチラベル分類に基づく列に対応するエンティティ推定

列ごとに，NE 列とリテラル列それぞれに対する列のエンティティを推定する手法を提案する．本手法では，列に対するエンティティ推定をマルチラベル分類 [Tsuoumakas 06] として定式化する．NE 列とリテラル列は異なる特徴を持つため，それぞれの特徴量を設計する．その特徴量を用いて，NE 列とリテラル列それぞれにマルチラベル分類器を学習する．学習したマルチラベル分類器を用いて，新たな列に対応するエンティティ推定を行う．ここでは，学習データおよび新たな列が NE 列とリテラル列のどちらかという分類は事前になされていると仮定する．

3.1 NE 列の特徴量

NE 列に対する特徴量は，各要素に対応づけられる可能性のあるエンティティを用いて定義する．定義より NE 列の各要素はエンティティに割り当てることが可能であり，それらによって特徴付けられる．知識ベースのベクトル空間への埋め込みを用いると，エンティティはベクトルで表現できる．ここでは，ユークリッド空間への埋め込みに比べて少ない次元でエンティティ間の構造を表現することが可能なポアンカレ埋め込み [Nickel 17] を用いて，知識ベースを埋め込む．NE 列の特徴量は，以下の手順で定める．(1) NE 列の各要素について，要素中の単語を同義語とするようなエンティティの集合を求め，それの和集合を列を特徴づける集合とする．(2) その集合の各要素について，知識ベースの埋め込みによって対応づけられたベクトルを求める．(3) 求めたベクトル集合について，それの平均と分散とをその NE 列の特徴量とする．

3.2 リテラル列の特徴量

リテラル列に対する特徴量は，Pham らの手法を参考にして定義する．リテラル列は，NE 列と異なり各要素がエンティティとは対応づけられない．従って，NE 列とは異なり，各要素の文字列の特徴を特徴量とする．また，リテラル列は数値や記号を含む文字列を要素とすることが多い．これは数値や記号はエンティティに対応づけることが難しいためであり，そのため日付や時刻，値段などが要素となる．リテラル列の 1 つの列を，文字列で表される要素の集合として解釈する．Pham らの手法は，2 つの列の間，つまり集合間に対して複数の類似度を用い，その類似度として数値的な類似度と文字列的な類似度を利用している．本提案手法では，類似度の代わりに，数値的な特徴量と文字列的な特徴量を用いる．数値的な特徴量は，列内の要素中の数値で構成される集合に対する最小値と最大値・中央値・平均・標準偏差・3 次モーメント・歪度・尖度とする．文字列的な特徴量は，今回の実験で用いたデータ中に頻出する記号である「:」,「,」,「.」,「/」の要素ごとの平均出現回数とする．

3.3 列に対応するエンティティ推定

本手法では，列の種類ごとにマルチラベル分類器を用いて，列に対応するエンティティを推定する．マルチラベル分類の最も簡単な手法として，Binary Relevance [Tsuoumakas 06] を用いる．Binary Relevance は，各ラベルごとに 2 クラス分類器を用意し，マルチラベル分類を実剣する方法である．この方法では，各分類器は入力に対してそのラベルが対応するかどうかを出力する．ここでは，ラベルをエンティティと読み替えて，マルチラベル分類器を学習する．また，2 クラス分類器にランダムフォレストを用いる [Breiman 01]．新たな入力列を特徴ベクトルに変換し，学習したマルチラベル分類器を用いて，各エンティティの出力確率を求める．これにより，新たに入力された列に対して，各エンティティがどの程度対応づけられるかを推定する．

4. 複数列に対する同時概念推定

列ごとに推定する手法を拡張して，テーブル中の複数列に対して同時に列のエンティティを推定する手法を提案する．テーブル中の各変数間の関係を表現するためにマルコフ確率場を用いる．提案手法では，テーブル中のリテラル列に対応するエンティティも同時に推定することができる．

4.1 テーブルの中の変数とポテンシャル関数

テーブルの行数を R とし，列数を C とする．r 行 c 列目の要素に対応する観測変数を x_{rc} とし，テーブルのタイトルに対応する潜在変数を t，テーブルの c 番目の列のエンティティに対応する潜在変数を y_c とする．観測変数は文字列を値とし，

潜在変数はエンティティを値とする．これらの変数を用いて，テーブルに対するマルコフ確率場を定義する．例えば，2 行 3 列のテーブルに対応するマルコフ確率場は図 1 のようになる．Limaye らの提案したモデルと異なり，観測変数 x_{rc} が直接 y_c に影響するモデルとなっている．

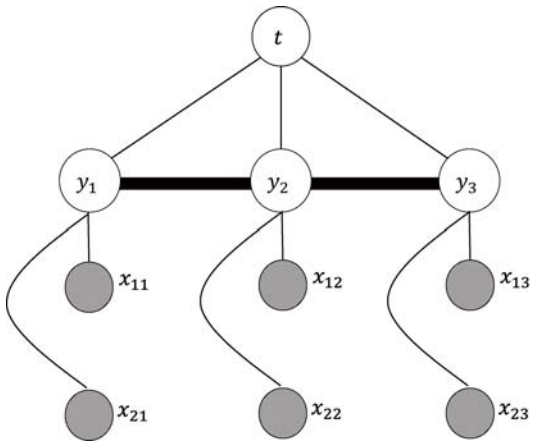

図 1: 2 行 3 列のテーブルに対するグラフィカルモデルの例．y_c を貫く太い線は y_c 間の全結合を表す．

各変数間のポテンシャル関数を以下で定義する．(1) 潜在変数 y_c と列 c の観測変数の集合 $\mathbf{x}_c = \{x_{rc}\}_{r=1}^{R}$ のポテンシャル関数 $\phi_{yx}(y_c, \mathbf{x}_c)$ は，学習データ中の $\mathbf{x}_c$ と似た列のエンティティが y_c であれば，大きな値をとる関数である．具体的には，$\phi_{yx}(y_c, \mathbf{x}_c) = f_{y_c}(\mathbf{x}_c)$ と定義する．ただし，$f_{y_c}(\mathbf{x}_c)$ は $\mathbf{x}_c$ を入力としたときのマルチラベル分類器の出力が y_c をラベルとして出力する確率値を表す．マルチラベル分類器によって ϕ_{yx} を定義することによって，学習データから対応づけ方が自動的に抽出される．

(2) 潜在変数 y_c と $\mathbf{y}_{-c}$ のポテンシャル関数 $\phi_{yy}(y_c, \mathbf{y}_{-c})$ は，y_c と $\mathbf{y}_{-c}$ が学習データ中で共起している場合大きな値をとる関数である．ここで $\mathbf{y}_{-c}$ は $\mathbf{y}$ から y_c を除いたものとする．このポテンシャル関数 ϕ_{yy} は，ϕ_{yx} と同様にマルチラベル分類器の出力を用いて，$\phi_{yy}(y_c, \mathbf{y}_{-c}) = g_{y_c}(\mathbf{y}_{-c})$ と定義される．$g_{y_c}(\mathbf{y}_{-c})$ は，$\mathbf{y}_{-c}$ を特徴ベクトルに変換したものを入力としたマルチラベル分類器の出力が，y_c をラベルとして出力するときの確率値を表す．知識ベースの埋め込みを用いて，以下の手順で $\mathbf{y}_{-c}$ を特徴ベクトルに変換する．$y_{c'} \in \mathbf{y}_{-c}$ に対応したベクトルを埋め込み空間から求める．得られたベクトルの集合を $\mathbf{V}_{-c}$ として，$\mathbf{V}_{-c}$ の平均と分散を $\mathbf{y}_{-c}$ の特徴量とする．学習データを用いて，このような特徴量を入力とし，c 列目に対応するエンティティ集合を出力とするマルチラベル分類器 $\mathbf{g}$ を学習する．

(3) 潜在変数 t と y_c のポテンシャル関数 $\phi_{ty}(t, y_c)$ は，y_c に対して意味的に近いタイトル t に対して大きな値をとる関数である．具体的には，ϕ_{ty} は埋め込み空間上での距離関数を $d(t, y_c)$ として，$\phi_{ty}(t, y_c) = \exp\{-wd(t, y_c)\}$ と定義される．ただし，w は重みパラメータである．

4.2 推論

これらのポテンシャル関数を用いて，マルコフ確率場上の同時確率は以下の式で定義される．

$$p(\mathbf{X}, \mathbf{y}, t) \propto \prod_{c=1}^{C} \phi_{yx}(y_c, \mathbf{x}_c)\phi_{yy}(y_c, \mathbf{y}_{-c})\phi_{ty}(t, y_c) \qquad (1)$$

観測変数の集合 $\mathbf{X} = \{x_{rc}\}$ が与えられた下で，式 1 を最大化する $\mathbf{y}$ を求める．

潜在変数はエンティティを値として取るため，式 1 を最大にする潜在変数の組を求めることは，組合せ最適化となり困難である．ここでは，初期値を適当に与え，ギブスサンプリングによってこの同時分布の最大化を近似する．具体的には，各列ごとに ϕ_{yx} の値が最大となる y_c を求めて初期値とする．それから，全ての y_c が固定された下で，$\phi_{ty}(t, y_c)$ が閾値以上になる全ての t を候補として擬似事後分布からサンプリングする．さらに，t と $\mathbf{y}_{-c}$ が固定された下で，y_c に対応するマルチラベル分類器の出力できる候補集合から，サンプリングする．これらを 300 回程度繰り返し行い，式 1 を最大にする $\mathbf{y}$ を求める．

5. 評価実験

5.1 実験設定

UCI Machine Learning Repository [Lichman 13] の 183 の CSV データを利用する．各 CSV データをテーブルとして，実験に用いる．これらのテーブルは，781 の NE 列と 4109 のリテラル列によって構成される．

知識ベースには，WordNet [Miller 95] を用いる．特に，列に対応するエンティティは名詞であることが多いことから，本実験ではエンティティを名詞に限る．WordNet を用いて UCI のデータセットを複数人によってアノテーションする．このアノテーション結果を真値とする．ただし，1 つの列に対して複数のエンティティが真値となることがある．

ポアンカレ埋め込み [Nickel 17] を用いて，WordNet のエンティティを上位下位概念の関係を保存するように 5 次元の双曲空間に埋め込む．ポアンカレ埋め込みを用いることで，ユークリッド空間にに比べて低次元であっても，うまく構造を保存する．

5.2 評価指標

MAP@k と nDCG@k・類似度に基づくスコア（類似度@k）を用いて評価する．MAP@k と nDCG@k は，ランキングに関する一般的な評価指標である [Clarke 08]．また，類似度@k は，真のエンティティ集合と推定結果がどれほど類似しているかを表現する．上位 k 位までの推定結果について，真のエンティティとの Wup 類似度 [Wu 94] を求め，その確信度に応じた重み付き平均で定義する．Wup 類似度は，0 から 1 の範囲の値をとり，2 つのエンティティが類似している場合大きな値をとる．本実験では $k = 5$ として評価する．

5.3 NE 列に対する推定

NE 列に対して，列に対応するエンティティ推定の実験結果を表 1 に示す．Limaye らの手法であった列間の関係を無視し，独立な列として列に対応するエンティティ推定を行う手法と，Pham らの手法の中で文字列的な類似度のみを用いた手法を比較手法とする．ただし，ここで用いる提案手法は，列独立に NE 列に対応するエンティティ推定を行う手法である．提案手法は，すべての指標で比較手法を上回る．

5.4 リテラル列に対する推定

リテラル列に対して，列に対応するエンティティ推定の実験結果を表 2 に示す．比較手法として，Pham らの類似度に基づくエンティティ推定手法を用いる．ただし，ここで用いる提案手法は，列独立にリテラル列に対応するエンティティ推定を行う手法である．提案手法は比較手法を上回る性能を示した．Pham らの手法と提案手法の違いを考えるために，学習データ中に，メートルとインチのように単位は異なるが列に対応す

表 1: NE 列に対する，列に対応するエンティティ推定の実験結果．Limaye らの手法に基づく比較手法（比較手法 1）と類似度に基づく比較手法（比較手法 2），マルチラベル分類による提案手法の比較．

手法	MAP@5	nDCG@5	類似度@5
比較手法 1	0.021	0.077	0.411
比較手法 2	0.196	0.328	0.470
提案手法	**0.314**	**0.365**	**0.526**

るエンティティが同じ場合を考察する．Pham らの手法では，この場合，類似度が低いにも関わらず同じエンティティを指すため，適切な重みを学習することが困難になる．一方で，提案手法は，この場合それぞれを特徴量に変換して考えるため，学習に大きな影響を与えることが少ないと考えられる．

表 2: リテラル列に対する，列に対応するエンティティ推定の実験結果．類似度に基づく比較手法とマルチラベル分類による提案手法の比較．

手法	MAP@5	nDCG@5	類似度@5
比較手法	0.270	0.330	0.467
提案手法	**0.397**	**0.430**	**0.590**

5.5 テーブル全体に対する推定

複数列のテーブルに対して，列に対応するエンティティ推定の実験結果を表 3 に示す．Limaye らの手法をリテラル列に対して拡張した手法と，列間の依存関係を無視した提案手法，さらに列間の関係を含んだ提案手法の 3 つの比較である．列間の関係を含んだ提案手法の評価は，提案手法において列独立に行う場合より優れている．これより，列間の関係をマルチラベル分類器を用いてモデル化し，タイトルを表現する潜在変数を導入したことによって，推定性能が向上するといえる．また，Limaye らの手法は，様々な類似度を用いてポテンシャル関数を定義するため，適切な類似度を設定する必要がある．提案手法は，類似度の決定の代わりに知識ベースの埋め込みと，マルチラベル分類器がポテンシャル関数を定義している．学習データがある環境では，それらに応じたポテンシャル関数を定義できる提案手法が優れていると考えられる．

表 3: 複数列のテーブルについて，列に対応するエンティティ推定を行った実験結果．Limaye らの手法を元にリテラル列に拡張した手法（比較手法）と，列間の依存関係を無視して実行した提案手法（列独立），列間の依存関係を考慮する提案手法（列依存）の 3 つの手法の比較．

手法	MAP@5	nDCG@5	類似度@5
比較手法	0.225	0.291	0.480
提案手法 (列独立)	0.351	0.413	0.537
提案手法 (列依存)	**0.464**	**0.741**	**0.635**

参考文献

[Breiman 01] Breiman, L. (2001). Random forests. *Machine learning*, 45(1), 5–32.

[Clarke 08] Clarke, C. L., M. Kolla, G. V. Cormack, O. Vechtomova, A. Ashkan, S. Büttcher, and I. MacKinnon (2008). Novelty and diversity in information retrieval evaluation. *Proceedings of the 31st annual international ACM SIGIR conference on Research and development in information retrieval*, 659–666. ACM.

[Deng 13] Deng, D., Y. Jiang, G. Li, J. Li, and C. Yu (2013). Scalable column concept determination for web tables using large knowledge bases. *Proceedings of the VLDB Endowment*, 6(13), 1606–1617.

[Lichman 13] Lichman, M. (2013). UCI Machine Learning Repository [http://archive.ics.uci.edu/ml]. Irvine, CA: University of California, School of Information and Computer Science.

[Limaye 10] Limaye, G., S. Sarawagi, and S. Chakrabarti (2010). Annotating and searching web tables using entities, types and relationships. *Proceedings of the VLDB Endowment*, 3, 1338–1347.

[Miller 95] Miller, G. A. (1995). WordNet: a lexical database for English. *Communications of the ACM*, 38(11), 39–41.

[Mulwad 13] Mulwad, V., T. Finin, and A. Joshi (2013). Semantic message passing for generating linked data from tables. *International Semantic Web Conference*, 363–378. Springer.

[Neumaier 16] Neumaier, S., J. Umbrich, J. X. Parreira, and A. Polleres (2016). Multi-level semantic labelling of numerical values. *International Semantic Web Conference*, 428–445. Springer.

[Nickel 17] Nickel, M., and D. Kiela (2017). Poincaré embeddings for learning hierarchical representations. *Advances in Neural Information Processing Systems 30*, 6338–6347.

[Pham 16] Pham, M., S. Alse, C. A. Knoblock, and P. Szekely (2016). Semantic labeling: a domain-independent approach. *International Semantic Web Conference*, 446–462. Springer.

[Tsuoumakas 06] Tsoumakas, G., and I. Katakis (2006). Multi-label classification: an overview. *International Journal of Data Warehousing and Mining*, 3(3), 1–13.

[Venetis 11] Venetis, P., A. Halevy, J. Madhavan, M. Paşca, W. Shen, F. Wu, G. Miao, and C. Wu (2011). Recovering semantics of tables on the web. *Proceedings of the VLDB Endowment*, 4(9), 528-538.

[Wu 94] Wu, Z., and M. Palmer (1994). Verbs semantics and lexical selection. *Proceesings of the 32nd Annual Meeting on Association for Computational Linguistics(ACL '94)*, 6, 133–138.

[Zhang 17] Zhang, Z. (2017). Effective and efficient semantic table interpretation using tableminer+. *Semantic Web*, 8(6), 921-957.