3D1-6

エンティティ・カーネル：複数関係ネットワークに対するリンクベース回帰手法

Entity Kernel: A Link-based Regression Method for Multi-relational Network

大澤 昇平[*1]
Shohei Ohsawa

松尾 豊[*1]
Yutaka Matsuo

[*1] 東京大学
The University of Tokyo

This paper addresses a classification/regression problem for named entities such as people, places, and products. Assuming that we want to classify whether a person is in a profession (e.g., researcher) or not, we must construct features to make the classification problem. Typically we use attributes of the entity itself (e.g., gender and age) as well as relations among entities (e.g., the number of friends in a profession). Typical link-based classification deals with single relations, but apparently there are multiple relations in the real world. This paper presents a domain-independent link-based classification/regression method for use with relations of multiple types. We construct a tensor structure that is equivalent to a given multi-relational entity network. It enables us to make an entity kernel that can be used for kernel-based methods. Evaluation showed that our approach achieved 8.1% root mean square error, which is better than any other method.

1. はじめに

本論文は人物、場所、製品を初めとする名前付きエンティティに対する回帰問題を扱う。たとえば、ある人物が研究者かそうではないかという問題を解くことを考える。この場合、ある識別関数に関する回帰問題を解くための素性を構築する必要がある。典型的な方法としては、エンティティそのもの（性別・年齢）に加え、エンティティ間の関係（彼が何人の研究者の友人を持っているか）を用いることが考えうる。ソーシャルメディア分析の分野においては、[Cheng 2010]はユーザの友人の場所を素性に加えることで、ユーザの場所の回帰の精度を改善することが報告されている。本論文では、エンティティ関係を用いた回帰手法を、**リンクベース回帰**と呼ぶ。

既存のリンクベース回帰手法は単一の関係を扱っているが、実世界においては複数の関係が存在している場合が多い。たとえば、社会心理学においては、人物間の関係は、友人関係に加え敵対関係があるとされる。また、同一種類のエンティティ間の関係に加え、人物と組織といった、異なる種類のエンティティ間の関係も存在する。さらに、セマンティックウェブの分野においては、様々な関係の種類を持ったオープンで大規模な複数関係ネットワークを利用することが可能になっている。たとえば、DBPedia [Auer 2007]と YAGO [Suchanek 2008]は、Wikipedia から自動的に抽出されたこのようなネットワークであり、有名人や場所、製品間の関係を持っている。

複数関係ネットワークを扱えるいくつかの回帰手法は過去に提案されている[Balmin 2004]ものの、これらはモデルパラメータとしてそれぞれの関係に重みを人間が与えしななくてはならず、これらを調整するための適切な方法についてはこれまで議論されて来なかった。このような既存手法においては、一般にモデルパラメータの精度は多峰性であり、最適化は凸問題にならない。加えて、パラメータの増加は組み合わせ爆発を引き起こし、調整は時間がかかるようになり、結果的に焼きなまし法や遺伝的アルゴリズムといったメタヒューリスティクスが必要になる。このような理由から、多種類のエンティティ関係をリンクベース回帰に用いることは手動で関係を選択することで行われており、結果的に多くの手法がドメイン依存になっていた。

本論文は、複数関係ネットワークに対するドメイン非依存なリンクベース回帰手法を提案する。本論文では、複数関係ネットワークに対して等価なテンソル構造を構築する。これはカーネルベースの手法に用いることができるエンティティ・カーネルを構成することを可能にする。エンティティ・カーネルは、与えられた複数関係ネットワークから得られるテンソル構造をガイドに二つのエンティティ間の類似度を演算するものであり、様々な機械学習の資産を活用することで回帰問題を解くことを可能にする。特に、サポートベクターマシン(SVM)にエンティティ・カーネルを適用することにより、各関係種類の重みを調整する問題を、ある陰な素性空間上における凸問題に還元する。

本論文では、エンティティ・カーネルを Facebook のファンページの《いいね！》数を予測する問題に適用する。Facebook ファンページは、各ファンページをエンティティとみなすことで、複数関係ネットワークをなす構造ということができる。評価結果では、本手法は他の手法に対して最良の結果である、8.1%の二乗平均平方根誤差を記録している。実験の結果抽出された重要な述語は、エンティティ間の相互影響ネットワークをなしており、各関係種類の重みに対して俯瞰的な視点を与えている。

第 2 章では、テンソルモデルの説明を行う。第 3 章では素性構築について説明を行う。第 4 章では、実験結果について述べる。関連研究と議論について述べた後、結論について述べる。

連絡先：大澤昇平、東京大学大学院工学系研究科、東京都文京区弥生 2-11-16 工学部 9 号館 210 ohsawa@weblab.t.u-tokyo.ac.jp

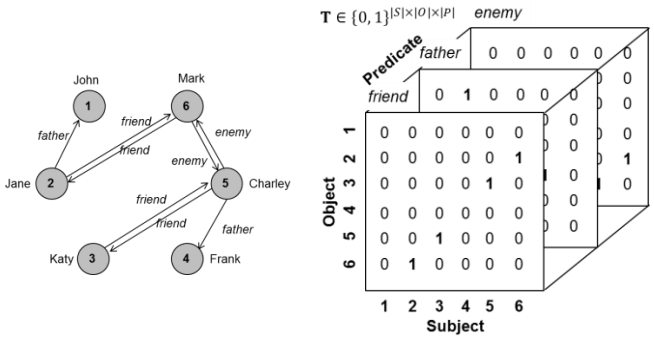
2. 前提知識

2.1 データモデル

上記の問題を解くために、本論文では、RDF[World 1999]ベースのテンソルモデルを定義する。図 1 は、テンソルモデルの概念図である。図 1(a)のRDFデータは 3 つの関係、**friend**, **father**, **enemy**を持っている。**friend**および**enemy**は逆の意味を持っているため、それぞれの関係を区別する必要がある。そこで、図 1(b)のようなテンソル構造 $\mathbf{T} \in \{0,1\}^{|S|\times|O|\times|P|}$を導入する。$S, P, O$はそれぞれ、すべての主語、述語、目的語の集合である。この 3 次元テンソルは要素$t_{spo} \in \{0,1\}$を持つ。これは、対応するRDFデータがトリプル(s,p,o)を持てば 1 を取り、それ以外は 0 を取る。

次に、属性ベクトル$\boldsymbol{x} \in \Re^{r+l}$を定義する。これは人気度などの要素の属性を表現する。ただし、rは全リソースの数であり、lは全リテラルの数である。

RDFテンソルに対し、いくつかの操作をすることで、意味的な側面を抽出することができる。s-**ファセット** $\mathbf{T}|_{s=A}$はテンソルを主語でスライスするファセットである。結果的に、$\mathbf{T}|_{s=A}$ はAに関連する関係をもつことになる。本論文では同様に、p-**ファセット**、o-**ファセット**を定義する。



(a) 複数関係ネットワーク　　(b) 対応するテンソル構造

図 1　データモデル

2.2 問題定義

定義 1　RDFテンソル **T**に対し、複数関係ネットワークに対するリンクベース回帰問題とは、以下によって定義される。

入力　RDFテンソル **T**、訓練マッピング $l_t: V_t \to \Re$
出力　推定された回帰関数 $\hat{l}: V \to \Re$.

3. 提案手法

本論文では、対象とする回帰問題を、テンソルベース回帰の副問題として解く。すなわち、与えられたエンティティに対し、回帰式 $\hat{x}_s = \boldsymbol{w}^{\mathrm{t}}(T_s\boldsymbol{x})$ を、予測値を得るために用いる。T_sはsのs-ファセットであり、$\boldsymbol{w} \in \Re^{|P|}$はすべての述語に対する重みベクトルである。最適な重みベクトルは、予測値と訓練値の間の誤差を最小化するものとして求められる。

3.1 リッジ回帰

リッジ回帰においては、目的関数 $J(\boldsymbol{w})$ は以下の式によって定義される。

$$J(\boldsymbol{w}) = \frac{1}{2}\sum_{i=1}^{|S|}\left(\boldsymbol{w}^t\left(T_{s_i}\boldsymbol{x}\right) - x_{s_i}\right)^2 + \frac{\lambda}{2}\boldsymbol{w}^{\mathrm{t}}\boldsymbol{w}.$$

この関数を$\boldsymbol{w}$に対して最小化することにより、$\tilde{\boldsymbol{w}} = \arg\min J(\boldsymbol{w}) = U^{\mathrm{t}}(UU^{\mathrm{t}} + \lambda I)^{-1}\boldsymbol{x}$を得る。ここで、$U$は$i$行目が$\left(T_{s_i}\boldsymbol{x}\right)^{\mathrm{t}}$であるような行列である。したがって、新たに到着したエンティティs_{new}に対して以下の式を用いることで、s_{new}の属性を求めることができる。

$$x_{s_{\mathrm{new}}} = \boldsymbol{x}^{\mathrm{t}}T_{s_{\mathrm{new}}}^{\mathrm{t}}U^{\mathrm{t}}(UU^{\mathrm{t}} + \lambda I)^{-1}\boldsymbol{x}.$$

3.2 エンティティ・カーネル

予測式$\hat{x}_s = \boldsymbol{w}^{\mathrm{t}}(T_s\boldsymbol{x})$を注意深く観察すると、$\boldsymbol{x}$に関する線形関数になっていることが分かる。したがって、この式に対してカーネルトリックを適用することができる。$T_s\boldsymbol{x}$を基底関数$\boldsymbol{\phi}$で置換することにより、$\hat{x}_s = \boldsymbol{w}^{\mathrm{t}}\boldsymbol{\phi}(T_s\boldsymbol{x})$を得る。ここで、リッジ回帰の手法をこの式に適用すると、目的関数は以下のように書くことができる。

$$J(\boldsymbol{w}) = \frac{1}{2}\sum_{i=1}^{|S|}\left(\boldsymbol{w}^{\mathrm{t}}\boldsymbol{\phi}\left(T_{s_i}\boldsymbol{x}\right) - x_{s_i}\right)^2 + \frac{\lambda}{2}\boldsymbol{w}^{\mathrm{t}}\boldsymbol{w}.$$

この式を$\boldsymbol{w}$に関して最小化することにより、$\tilde{\boldsymbol{w}} = \arg\min J(\boldsymbol{w}) = \Phi^{\mathrm{t}}(\Phi\Phi^{\mathrm{t}} + \lambda I)^{-1}\boldsymbol{x}$を得る。ここで、$\Phi$は$i$番目の行が$\boldsymbol{\phi}\left(\boldsymbol{T}_{\boldsymbol{s}_i}\boldsymbol{x}\right)^{\mathrm{t}}$であるような行列である。したがって、新たに到着したエンティティs_{new}に対し、以下の式により対応する属性を予測することができる。

$$\hat{x}_{s_{new}} = \boldsymbol{\phi}\left(T_{s_{new}}\boldsymbol{x}\right)^{\mathrm{t}}\Phi^{\mathrm{t}}(\Phi\Phi^{\mathrm{t}} + \lambda I)^{-1}\boldsymbol{x}..$$

ここで、カーネル関数$k(\boldsymbol{x}, \boldsymbol{x}') = \boldsymbol{\phi}(\boldsymbol{x})^{\mathrm{t}}\boldsymbol{\phi}(\boldsymbol{x})$を用いることで、式を以下のように書くことができる。

$$\hat{x}_{s_{\mathrm{new}}} = \mathbf{k}\left(T_{s_{\mathrm{new}}}\boldsymbol{x}\right)^{\mathrm{t}}(K + \lambda I)^{-1}\boldsymbol{x}.$$

ただし、

$$\mathbf{k}\left(T_{s_{\mathrm{new}}}\boldsymbol{x}\right) = \begin{pmatrix} k(T_{s_1}\boldsymbol{x},\ T_{s_{\mathrm{new}}}\boldsymbol{x}) \\ k(T_{s_2}\boldsymbol{x},\ T_{s_{\mathrm{new}}}\boldsymbol{x}) \\ \vdots \\ k(T_{s_{|S|}}\boldsymbol{x},\ T_{s_{\mathrm{new}}}\boldsymbol{x}) \end{pmatrix},$$

$$K = \begin{pmatrix} k(T_{s_1}\boldsymbol{x}, T_{s_1}\boldsymbol{x}) & \cdots & k(T_{s_1}\boldsymbol{x}, T_{s_{|S|}}\boldsymbol{x}) \\ \vdots & \ddots & \vdots \\ k(T_{s_{|S|}}\boldsymbol{x}, T_{s_1}\boldsymbol{x}) & \cdots & k(T_{s_{|S|}}\boldsymbol{x}, T_{s_{|S|}}\boldsymbol{x}) \end{pmatrix}.$$

この式は、予測値がカーネル関数のみによって表現されることを表している。このカーネル関数は、$\{s_i\}$のみによって制御される。

上記を踏まえ、エンティティ・カーネルを以下のように定義する。

定義 2　与えられたエンティティs_1, s_2、$k \in \mathcal{F}$ ($\mathcal{F}$ は再生核ヒルベルト空間)、RDFテンソル$\mathbf{T}$、属性ベクトル$\boldsymbol{x}$に対し、**エンティティ・カーネル**は次式によって与えられる。

$$k_e(s_1, s_2; k, \mathbf{T}, \boldsymbol{x}) \stackrel{\text{def}}{\equiv} k\left(T_{s_1}\boldsymbol{x}, T_{s_2}\boldsymbol{x}\right).$$

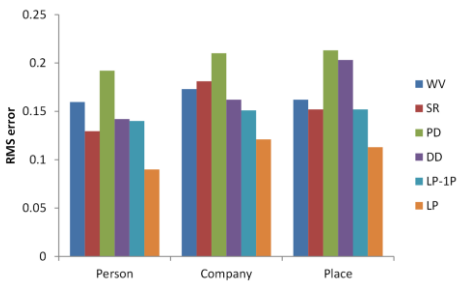


図 2　予測手法の比較

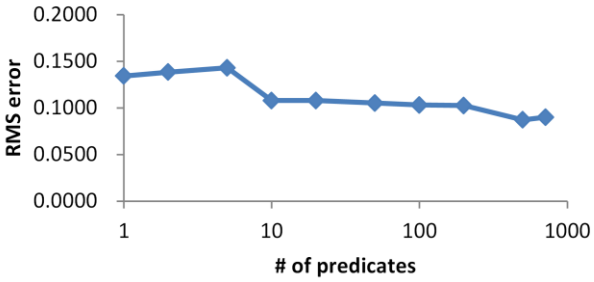


図 3　述語数を変化させた場合の精度

エンティティ・カーネルはSVMや主成分分析など、様々なカーネルベースの手法に対して用いることができる。エンティティ・カーネルはRDFデータを生かし、$\mathbf{T}$と$\boldsymbol{x}$のみに依存する。$\mathbf{T}$は一般に疎行列であることから、テンソル分解によってエンティティ・カーネルの値を求める計算量を削減できることが期待できる。

以上を踏まえ、本論文における回帰問題を、エンティティ・カーネルとSVMを用いて解く。

4. 実験

本章では、エンティティ・カーネルの回帰問題に対する性能を評価する。

4.1 Facebook Like データセット

エンティティ・カーネルの回帰問題に対する性能を評価するため、Facebook Likeデータセットを用いる。これはFacebookから独自にクロールされたデータである。このデータセットは、いくつかのエンティティと、それらの間の関係性を保持している。それぞれのエンティティは、Facebookページに対応している。

4.2 複数関係ネットワークに対するリンクベース回帰手法の性能

まず、回帰問題に対する評価を行う。データセットとしては、FacebookページとDBPediaを統合したRDFデータを用いる。4 分割交差検定を行い、評価指標としてそれぞれの検定における二乗平均平方根(RMS)誤差の平均を用いる。本実験では 3 つのデータセット、`Person`、`Company`、`Place`を用いる。それぞれのデータセットは 500,000 ファンページで構成され、各ファンページはデータセット名と同一のクラスに属するDBPediaのエンティティと関連付けられている。リンクベース回帰の有用性について検証するため、本手法を以下の 4 つのベースラインと比較する。

Wikipedia閲覧数(WV) 与えられたエンティティのスコアを、同一の名前を持つWikipedia記事の閲覧数に定数を乗ずることによって推定する。乗数は最小二乗法によって推定する。

検索結果数(SR) 与えられたエンティティのスコアを、キーワードとしてタイトルを検索し、得られた検索結果数によって推定する。対象検索エンジンとしてはBingを選んだ。

ファンページ説明文(PD) ファンページの説明文を素性として用いる。本論文ではbag-of-wordsモデルを採用し、テキストを各単語のTF-IDF値に基づくベクトル空間に写像する。すなわち、次元数は単語数に対応し、各次元の値はTF-ID値に相当する。

DBPedia説明文(DD) この手法はDBPediaのショート・アブストラクトを素性として用いる。ベクトル空間の構成方法はPDと同一である。

単一述語によるリンクベース回帰(LP-1P) この手法はほとんど提案手法と同じであるが、述語の属性を考慮しない点が異なる。この手法は様々なエンティティ関係を単一の関係とみなす。本論文ではこの手法を述語の多様性が重要であることを確認するために実験する。

図 2 は、いくつかの予測手法の性能を示している。このグラフを観察すると、本手法がRMS誤差の観点から優れていることがわかる。RMS誤差は、Personデータセットにおいて WVの 56.3%となっている。SRはWVよりは良い結果となっている。LP-1Pはあまり良い性能を出していない。この実験結果は、複数のエンティティ関係を考慮することが重要であることを示している。

次に、RMS誤差の低減に述語の多様性がどのように貢献しているかを確認するため、回帰問題を述語の種類数を変化させることでどのように変化するかを調べる。図 3 は述語の種類数を変化させた場合の性能を示している。本論文ではRDFを用いているため、簡単に述語を増やすことができる。これは、GeoNamesやFalconのような他のデータソースと接続することで、性能を向上させることができることを示している。

5. 関連研究

リンクベース回帰　いくつかの既存研究が、エンティティ周辺の関係から与えられたエンティティに関する属性を予測する問題を扱っている。[Cheng *et al.*, 2010] は、Twitterユーザの位置情報をソーシャル・ネットワークに基づき予測している。実験的に得られた結果は、ソーシャルネットワークを考慮することで、精度を向上できることを示している。[Lu and Getoor, 2003]は、エンティティの関係を分類問題を解くのに採用している。彼らは**リンクベース分類**を提案し、与えられたエンティティのクラスを隣接エンティティのクラスの関数として推定している。この手法と類似して、[Hu *et al.*, 2009] はWikipediaのエンティティ関係を情報検索におけるユーザクエリ意図の分類問題に適用している。彼らは最初に少数のシード・エンティティからなる教師データを作成し、各エンティティのクラスを周辺に伝搬させる。こうした手法とは対照的に、本手法は述語によって区別される複数の関係を区別している点が異なる。実験結果は、述語を導入することが、モデルの制度改善に貢献していることを示している。

オントロジからの特徴量構築　本論文では、DBPediaからの自動的な特徴量構築について述べた。[Paulheim *et al.*

2012]はLinked Open Dataからの特徴量構築手法であるFeGeLODを提案している。これは自動的にSPARQLを生成し、RDFデータに対して特徴量を得るために適用する。しかし、この手法は同じ主語と述語に対して複数の目的語が存在する場合を考慮に入れていない。このケースは、DBPediaから特徴量を構築する場合に起こりうる。たとえば、db:parentは 2 つの目的語を取りうる。本論文では、複数の目的語が存在する場合について説明し、統合する手法について提案した。

6. 結論

本論文ではRDFのような複数関係ネットワーク上における回帰問題のための、エンティティ・カーネルについて提案した。最初に、扱う回帰問題の形式的な定義を与え、それに対する手法を提案した。実験結果は本手法が対象とする回帰問題を 8.1%のRMS誤差で解くことができることを示した。

本手法はFacebookのデータセットのみで実験しているが、本手法はRDFとエンティティ・カーネルの柔軟性によりTwitterのような他のドメインにも広く適用することが可能である。たとえば、Twitterにおいてフォロワー数の予測を行うには、ユーザアカウントを名前と説明文に基づきRDFと関連付ければよい。この場合においては、いいね数はフォロワー数に対応する。本論文におけるエンティティ・リンキング手法はシンプルであるため、既存手法を改善することも可能になる。

属性ベクトルに基づき効率的なエンティティ関係を求めることは、新しい研究領域を開拓することが期待できる。これは意味ネットワークの情報量を削減し、効率的な推論アルゴリズムの開発につながると考えられる。

参考文献

[Auer *et al.*, 2007] S. Auer, C. Bizer, G. Kobilarov, J. Lehmann, R. Cyganiak and Z. Ives. DBpedia: A Nucleus for a Web of Open Data. *The Semantic WebLecture Notes in Computer Science*, vol. 4825, 722-735, 2007.

[Balmin *et al.*, 2004] A. Balmin, V. Hristidis and Y. Papakonstantinou, ObjectRank: authority-based keyword search in databases *In Proceedings of the Thirtieth international conference on Very large data bases (VLDB2004)*, 2004

[Cheng *et al.*, 2010] Z. Cheng, J. Caverlee and K. Lee. You Are Where You Tweet: A Content-Based Approach to Geo-locating Twitter Users. *In Proceedings of the 19th ACM international conference on Information and Knowledge Management (CIKM2010)*, 2010

[Danescu *et al*., 2008] C. Danescu, G. Kossinets, J. Kleinberg and L. Lee. How opinions are received by online communities: a case study of amazon.com helpfulness votes. In *Proceedings of the 18th International Conference on World wide web (WWW2008)*, 2008

[Hu *et al.*, 2009] J. Hu, G. Wang, F. Lochovsky, J. T. Sun and Z. Chen. Understanding User's Query Intent with Wikipedia. In *Proceedings of the 19th International World Wide Web Conference (WWW2009)*, 2009.

[Kwak *et al*., 2010] H. Kwak, C. Lee, H. Park and S, Moon. What is Twitter, a social network or a news media?. In *Proceedings of the 19th International World Wide Web Conference (WWW2010)*, 2010.

[Lair *et al*., 2005] D. J. Lair, K. Sullivan and G. Cheney. Marketization and the Recasting of the Professional Self. *Management Communication Quarterly February*, vol. 18 no. 3 307-343, doi: 10.1177/0893318904270744, 2005.

[Lu and Getoor, 2003] Q. Lu and L. Getoor. Link-based Classification. In *Proceedings of 20th International Conference on Machine Learning (ICML2003)*, 2003

[Mendes *et al.*, 2010] P. N. Mendes, A. Passant and P. Kapanipath. Twarql: Tapping Into the Wisdom of the Crowd. In *Proceedings of the Sixth International Conference on Semantic Systems (I-SEMANTICS 2010)*, 2010.

[Milne *et al.*, 2007] D. Milne, I. H. Witten and D. M. Nichol. A Knowledge-Based Search Engine Powered by DBPedia. In *Proceedings of the Sixteenth ACM Conference on Conference on Information and Knowledge Management (CIKM 2007)*, 2007.

[Okazaki and Matsuo, 2008] M. Okazaki and Y. Matsuo. Semantic Twitter: Analyzing Tweets for Real-Time Event Notification. In *Proceedings of the 2008/2009 International Conference on Social Software: Recent Trends and Developments in Social Software (BlogTalk)*, 2008.

[Overell *et al.*, 2009] S. Overell, B. Sigurbjörnsson and R. van Zwol. Classifying tags using open content resources. In *Proceedings of the Second ACM International Conference on Web Search and Data Mining (WSDM 2009)*, 2009.

[Paulheim and Fürnkranz, 2012] H. Paulheim and J. Fürnkranz. Unsupervised generation of data mining features from linked open data. In *Proceedings of the Second International Conference on Web Intelligence, Mining and Semantics (WIMS2012)*, 2012

[Shen *et al.*, 2012] W. Shen, J. Wang, P. Luo and M. Wang. LINDEN: Lining Named Entities with Knowledge Base via Semantic Knowledge. In *Proceedings of the 21st International World Wide Web Conference (WWW2012)*, 2012.

[Suchanek *et al.*, 2007] F. M. Suchanek, G. Kasneci and G. Weikum. Yago: a core of semantic knowledge. In *Proceedings of the 16th International Conference on World Wide Web (WWW 2007)*, 2007.

[Vercoustre *et al.*, 2008] A. M. Vercoustre, J. A. Thom and J. Pehcevski. Entity ranking in DBPedia. In *Proceedings of the 2008 ACM Symposium on Applied Computing (SAC 2008)*. 2008.

[World 1999] World Wide Web Consortium. Resource Description Framework (RDF) Model and Syntax Specification, 1999.