2B1-1

ニューラルネットワークを用いた文書類似度の推定

Document similarity estimation using neural network

柳本 豪一*1
Hidekazu Yanagimoto

*1大阪府立大学
Osaka Prefecture University

It is important to estimate document similarity for text classification and information retrieval and so on. You should represent documents as feature vectors representing its content to get a good document similarity. In this paper I propose a document similarity estimation method using a neural network. To train a neural network with unsupervised learning fashion the neural network consists of multiple Restricted Boltzmann Machines. Evaluating the proposed method using stock market news, I confirmed that the proposed method can estimate document similarity according to their contents.

1. はじめに

テキスト分類や情報検索において文書間の類似度推定は重要である。自然言語処理では Bag-of-words モデルにより文書をベクトルで表現し、類似度が計算される。例えば、tf*idf では単語の出現頻度と文書頻度の逆数よりベクトルを求め、コサイン類似度が類似度として用いられる。また、確率モデルを用いた Okapi BM25[Robertson 94] などが提案されている。これらは、単語の出現頻度という単純な特徴量を修正し、高精度な類似度の計算を目指している。

画像認識の分野では SHIFT[Lowe 04] が画像の画素情報から新しい特徴量を作成して画像をベクトルで表現し、類似度が計算されている。また、最近ではニューラルネットワークの応用である Deep Learning[Bengio 09] を用いて画素データを変換し、高い認識率を実現する識別器を実現している。

本論文では、ニューラルネットワークを用いて単語の出現頻度で定義された特徴ベクトルを変換し、文書間の類似度の改善を目指す。具体的には大量の文書でトレーニングされた Restricted Boltmann Machine[Hinton 06] を組み合わせ、入力された特徴ベクトルを低次元の特徴ベクトルで表現する。そして、最終的に得られた特徴ベクトルを用いて文書間の類似度を計算する。評価実験では、株式ニュースを実験データとして利用し、株価の変動の観点から同一の影響を及ぼす記事間の類似度について検討を行う。これにより、ニューラルネットワークを用いた特徴ベクトルの変換により類似度が改善でき、関連性のある文書を見つけることができると確認できた。

2. 提案手法

本手法では、図 1 に示す多層ニューラルネットワークを用いて特徴量の変換を行う。ニューラルネットワークを用いて適切な特徴量の変換を行うためには、ニューラルネットワークを学習する必要がある。しかし、特徴量の変換を目的としているため、ニューラルネットワークを学習するための教師データを用意する事が困難である。また、中間層を多く含むニューラルネットワークを学習する際には、vanishing gradient problem[Hochreiter 98] が発生することが知られている。よって、通常の誤差逆伝搬法によって学習する事は困難である。

連絡先: 柳本 豪一，大阪府立大学工学研究科 電子・情報系専攻 知能情報工学分野，大阪府堺市中区学園町 1-1，TEL&FAX:072-254-9279，hidekazu@cs.osakafu-u.ac.jp

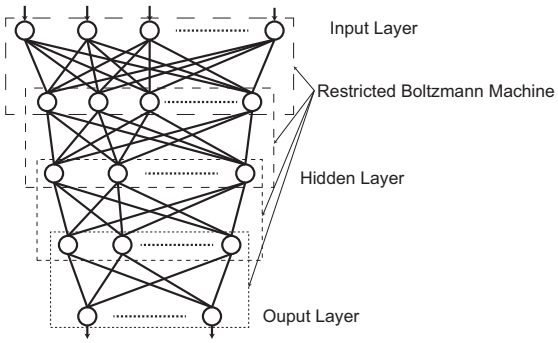

図 1: ニューラルネットワークの構成

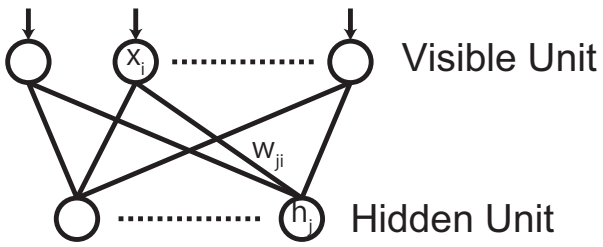

図 2: Restricted Boltzmann Machine の構成

そこで、この多層ニューラルネットワークを複数の Restricted Boltzmann Machine を接続したものとして見なす。Restricted Boltzmann Machine を図 2 に示す。Restricted Boltzmann Machine は学習において正解ラベルが不要なため、正解となる変換後の特徴ベクトルを用意する事なく、また vanishin gradient problem も回避できる。以下では、図 2 を用いて学習アルゴリズムの一つである Contrastive Divergence-1(CD-1) について説明する。詳細は Hinton らの論文 [Hinton 06] か機械学習の書籍 [Murphy 12] を参照してもらいたい。

まず、Restricted Boltzmann Machine における観測可能な入力層のニューロンと隠れ層のニューロンの状態に対する同時確率分布を定義し、その同時確率分布の最大値となるように重

Initialize W randomly
for each epoch do
  for each data $\mathbf{x}_i$ of size D do
    Compute $\mu_i = E[\mathbf{h}|\mathbf{x}_i, W]$
    Sample $\mathbf{h}_i \sim p(\mathbf{h}|\mathbf{x}_i, W)$
    Sample $\mathbf{x}_i^{'} \sim p(\mathbf{x}|\mathbf{h}_i, W)$
    Compute $\mu_i^{'} = E[\mathbf{h}|\mathbf{x}_i^{'}, W]$
    Accumlate $\mathbf{g} = \mathbf{g} + \mathbf{x}_i\mu_i^{\mathrm{T}} - \mathbf{x}_i^{'}\mu_i^{'\mathrm{T}}$
  Update parameters $W = W + \frac{\alpha}{D}\mathbf{g}$

図 3: CD-1 アルゴリズムの擬似コード

みを調整する。しかし、この Restricted Boltzmann Machine には観測不可能な隠れ層が含まれており、解析的に解く事が困難である。よって、同時確率の重みに対する勾配の近似を用いるものが CD-1 アルゴリズムである。

それでは、上の手順に沿ってアルゴリズムを導出する。まず同時確率分布を求める。このとき、Restricted Boltzmann Machine は層間のニューロンのみ接続され、層内のニューロンは接続されていない。このため、他層のニューロンの状態が決まっているときは、層内のニューロン同士は独立に状態を決定できる点に注意する。まず、Boltzmann Machine でよく考えられるエネルギー Energy$(\mathbf{x}, \mathbf{h})$ を導入する。

$$\mathrm{Energy}(\mathbf{x}, \mathbf{h}) = -\mathbf{b}^{\mathrm{T}}\mathbf{x} - \mathbf{c}^{\mathrm{T}}\mathbf{h} - \mathbf{h}^{\mathrm{T}}W\mathbf{x} \tag{1}$$

$\mathbf{b}$ と $\mathbf{c}$ は入力層および隠れ層の状態に対する重み、W は層間の接続の重みである。隠れ層の状態を全て考慮するために積分消去し、変数として明示的に表れない。

$$P(\mathbf{x}) = \frac{e^{-\sum_{\mathbf{h}} \mathrm{Energy}(\mathbf{x},\mathbf{h})}}{Z} \tag{2}$$

ここで Z は正規化項である。

上記の確率分布の最大値となる W を求める方法について説明する。これを求めるためには、$\frac{\partial \log P(\mathbf{x})}{\partial W}$ を求める必要があるが、解析的に求める事は一般的に困難である。CD-1 では初期状態と定常状態の Kullback-Leibler Divergence と初期状態から 1 ステップ進めた状態と定常状態の Kullback-Leibler Divergence との差により、勾配を近似する。これより、問題は初期状態から 1 ステップ進んだ状態を推定する事となるが、これはサンプリングにより解決する。以上の手法により求められる CD-1 アルゴリズムの擬似コードを図 3 に示す。

3. 実験

投資家に配信されている 2010 年度の株式ニュース T&C ニュースを用い、記事が及ぼす株価の変動の観点 (極性) が同じ記事同士の類似度について検討する。問題設定はセンチメント解析に近いため、素性としては形容詞、サ変動詞語幹を用いる。これは我々の従来研究において、株式ニュースの極性推定に有効な素性として判断したものである。

10,000 件の極性が不明な記事と 71 件の極性が判明している記事を用意し、ニューラルネットワークを学習する。71 件の記事は専門家によって極性を決めている。学習後に得られたニューラルネットワークで 71 件の記事の特徴ベクトルを変換することで得られたベクトルにより類似度を求め、同一極性の記事の類似度が高くなったかを調べる。

実験に用いる 10,071 件の記事から 2,604 個の素性を抽出し、これを 4 つの Restricted Boltzmann Machine により、1,000 次元、500 次元、250 次元へと低次元化し、最終的に 100 次元の特徴ベクトルとして表現する。

表 1: 提案手法による類似度を用いた関連記事数の変化

増加	減少	変化なし
36	17	18

表 1 に変換前と変換後より得られた特徴ベクトルを用いた時の類似度の上位 10 件の記事で極性が一致している記事数の変化についてまとめる。この結果から分かるように、学習したニューラルネットワークを用いて特徴ベクトルを用いることで、同一極性の記事の類似度が高くなることが分かった。また、特徴ベクトルを低次元化することにより適切な類似度推定が行なえているため、ニューラルネットワークの学習により関連した単語を効率的にまとめていると考えられる。この結果については構成されたニューラルネットワークを検証する事で、どのように単語がまとめられているかについて更なる検討が必要である。

4. おわりに

ニューラルネットワークにより特徴ベクトルを変換し、類似度を改善する方法を提案した。実験により、株価への影響が似ている記事間の類似度が高くなる特徴ベクトルを制裁される事が分かった。

今後は特徴ベクトルがどのように変換されているかについて調べる。また、テキスト分類などの手法に応用し、性能の評価を行う必要がある。

参考文献

[Robertson 94] Robertson, S. E., Walker, S., Jone, K. S., Hancock-Beaulieu, M., and Gatford, M.:Okapi at TREC-3, Proceedings of the Third Text REtrieval Conference, (1994).

[Lowe 04] Lowe, D. G.: Distinctive Image Features from Scale-Invariant Keypoints, International Journal of Computer Vision, Vol.60, No.2, pp.91–110, (2004)

[Bengio 09] Bengio. Y.: Learning Deep Architecture for AI., Foundations and Trends in Machine Learning, Vol.2, NO.1, pp.1-127, (2009).

[Hinton 06] Hinton, G. E. and Salakhutdinov, R. R: Reducing the Dimensionality of Data with Neural Network, SCIENCE, Vol.312, pp.504–507, (2006).

[Hochreiter 98] Hochreiter, G.: The Vanishing Gradient Problem During Learning Recurrent Neural Nets and Problem Solution, International Journal of Uncertainty Fuziness and Knowledge-Based Systems, Vol.6, No.2, pp.107–116, (1998).

[Hinton 06] Hinton, G. E., Osindero, S. T, and Tee, Y. W.:A Fast Learning Algorithm for Deep Blief Nets, Neural Computation, Vol.18, pp.1527–1554, (2006).

[Murphy 12] Murphy, K. P.: Machine Learning: A Probabilistic Perspective, The MIT Press, (2012).