

深層学習を用いたペアワイズ分類

古堂和音* 佐藤晴彦 小山聡 栗原正仁
(北大情報科学)[†] (北大工)

1 はじめに

ウェブ時代には正体のわからない同姓同名の人物の同一性のように、アイデンティティの推定が重要である。こういった推定を対象とした分類手法にペアワイズ分類がある。ペアワイズ分類は与えられた二つの例が同じクラスか否かを判別する分類問題である。従来、固定もしくは人手で調整されたペア間の類似度が研究されていたが、近年は機械学習を使った方法が研究されている。そのうちのひとつに、小山らによる研究がある。先行研究では、SVM とカーネル法により、異なる例からの異なる素性の組み合わせ情報から、素性間の関係性を q した分類器の学習法が示されている。一方で、SVM は、予め学習に使用する全データが必要である。データベースやウェブページは日々更新されるため、情報が追加されるたびに、過去の情報と合わせて学習をやり直すことは望ましい方法ではない。

そこで、本研究では、深層学習を用いたペアワイズ分類を提案する。深層学習は、大規模ニューラルネットワークである。深層学習は、ニューラルネットワークと同様にオンライン学習法をとれるため、日々増え続ける情報を使って分類器を追加的に学習可能である。また深層学習は、近年では各分野で高い性能を持つ学習器として注目を集めている。

本研究では、深層学習器の一つである Stacked Autoencoder とニューラルネットワーク、先行研究で用いられた SVM との比較によって、ペアワイズ分類における深層学習の有効性を測る。

2 研究目的

本研究の目的は、ペアワイズ分類におけるディープラーニングの有効性を検証することである。ディープラーニングの手法の一つである Stacked Autoencoder と、先行研究で使用された SVM との性能比較をし、ペアワイズ分類におけるディープラーニングの有効性について考察を加える。

3 関連研究

3.1 ペアワイズ分類

ペアワイズ分類は、与えられた二つの例が同じクラスに属するか否かを判定する分類である。エンティティのマッチングや重複発見は、データベースのコミュニティ

で長い間研究されてきたが、近年はこれを機械学習を用いて行う方法が考えられている。

3.2 問題の定式化

ここではペアワイズ分類の定義と定式化を行う。二つの例の組 x^α と x^β からなる新たな素性をペアインスタンスと呼ぶ。ペアワイズ分類は、このペアインスタンスが同じクラスに属するか否かを判別する問題と定義する。これは以下のように定式化できる。

$$f(x^\alpha, x^\beta) = \begin{cases} 1 & x^\alpha, x^\beta \text{ が同クラスの時} \\ -1 & \text{otherwise} \end{cases} \quad (1)$$

また、ペアワイズ分類と類似度との関連は深い。分類器の出力を $f(x^\alpha, x^\beta) \in [0, 1]$ のような連続値とすると、類似度を出力する関数と考えることもできる。この類似度に対して適切な閾値を設定してやれば、二値の分類器として使える。多くの分類器も、連続値で出力可能なため、これを類似度と考えることもできる。

4 深層学習

深層学習について説明する。深層学習は枠組みであり、実際の学習器の代表的なモデルとしては、Autoencoder 型、Deep Belief Net 型、Convolutional Neural Network 型が存在する。今回使用したモデルの Stacked Denoising Autoencoder と関連する要素技術を紹介する。

4.1 事前学習

多くの深層学習の手法で採用されており、深層学習の性能向上の要因の一つとなっているのが事前学習と呼ばれる学習方法である。ネットワークが 3 層以上の多層になる場合に、バックプロパゲーションに先行して行われる教師なし学習である。事前学習の目的は、教師あり学習に先立ってこの学習を行うことで、後のネットワークの学習が効率的に行われるためのネットワークの初期値を得ることである。

4.2 Autoencoder

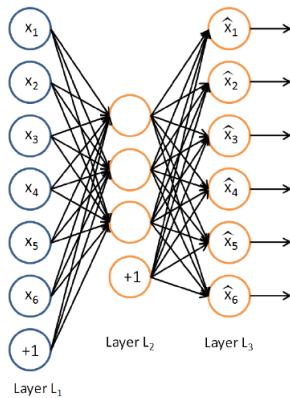


Fig. 1 Autoencoder の模式図

Autoencoder は入力層, 中間層, 出力層からなる 3 層のニューラルネットワークである. 出力層の数は入力層と等しくなるように制限される. Autoencoder は入力を中間層でエンコードし, 出力層でデコードする. Autoencoder は, エンコードとデコードによって入力を再構築するように学習を行う. エンコード, デコード, 再構築は以下のように示される.

$$\begin{aligned}
 \text{Encode} : f(\mathbf{x}) &= \sigma(\mathbf{w}_1\mathbf{x} + \mathbf{b}_1) \\
 \text{Decode} : g(\mathbf{y}) &= \sigma(\mathbf{w}_2\mathbf{y} + \mathbf{b}_2) \\
 \text{Reconstructed} : \mathbf{f}_n &= g(f(\mathbf{x}))
 \end{aligned} \tag{2}$$

学習は以下に示す再構築誤差を最小化するように行われる.

$$\begin{aligned}
 \text{Reconstruction Error} : \\
 L(\mathbf{x}, \mathbf{z}) &= \sum_j x_j \log z_j + (1 - x_j) \log(1 - z_j) \tag{3}
 \end{aligned}$$

図 4.2 のように中間層のユニット数を入力層のものより少なくすると, 次元削減を伴う特徴抽出を行うことができる. Autoencoder を多層にした場合, このネットワークは非線形の PCA と等価である.

Autoencoder の学習には, 合わせて様々な正則化が行われることが多い. 例えば, Sparse Autoencoder は再構築誤差に Sparse 項を導入することで, 中間層の活性のスパース性を誘導する. 他にも入力にノイズを付加し, それを除去するよう学習を行うことでより良い特徴を得ようとする Denoising と呼ばれる手法もよく行われる.

4.3 Stacked Autoencoder

Stacked Autoencoder(SA) は深層学習のモデルの一つである. Stacked Autoencoder の学習は事前学習とその後のファインチューニングに分けられる. Stacked Autoencoder は事前学習を Autoencoder を用いて行う. 事

前学習ではネットワークは層ごとに学習される. 例として, Stacked Autoencoder の最初の二層を学習することを考える.

このとき事前学習に用いるネットワーク構造は, Stacked Autoencoder の最初の二層に入力層と同じ数の活性化ユニットを持つ出力層を加えたものである. このネットワークを Autoencoder として学習し, 得られた入力層-中間層間の重みのみを, Stacked Autoencoder の最初の 2 層間の重みとする. 上層の学習を行うときは, 途中まで出来上がっている Stacked Autoencoder からの出力を Autoencoder の学習とする. こうして上層までネットワークを構築した後に, ファインチューニングに移る. ファインチューニングではバックプロパゲーションを用いてネットワーク全体の重みの微調整を行う教師あり学習である.

5 実験

6 実験目的

深層学習の一手法である Stacked Autoencoder を使ってペアワイズ分類に対する有効性を検証する. ペアインスタンスとその正解を訓練データとし, 訓練後の分類結果を確認する. SVM とニューラルネットワーク, そして, 本研究の提案手法である深層学習をそれぞれ比較する.

6.1 実験データ

実験データは DBLP[1] からの引用データセットを使った. このデータは, 論文 ID, 素性の番号, 素性の値という形式で得られる. 今回の実験では, このデータセットの中から同姓同名で混同されている数名を選び実験した.

6.1.1 ペアインスタンスの作成

ペアインスタンスの作成は, 以下の方法で行った. 母集団を 2 等分し, それぞれを訓練データの母集団, テストデータと検証データの母集団とする. そして, 各母集団の中で, 可能なすべての組み合わせを作り, ペアインスタンスとした. ペアインスタンスの作成法は, すべての方法で共通で, それぞれの例を順につなげたものとした. また, 学習結果がペアインスタンスの構成順序に依存しないように, ペアインスタンスの構成順序を逆にしたものもペアインスタンスとした. 各データセットの大きさや, そのデータ中に現れる素性数を表 1 に示す.

Table 1 Dataset Size

Abbreviated Name	Num of Papers	Training Data	Test Data	Validation Data	素性数
Gupta.A	398	39402	9850	9851	3444
Sato.M	157	12246	3061	3062	1724
Smith.J	389	75466	18866	18867	3652
Tanaka.K	176	15400	3850	3850	1832

6.2 実験設定

深層学習は入力層，中間層3層，出力層の全5層からなる Stacked Autoencoder を使った．それぞれの中間層の層数は入力層に近い側から，100-100-50 ユニットとした．比較のためニューラルネットワークも，Deep Learning Toolox を使って実装した．こちらの中間層は100ユニットとした．また，深層学習もニューラルネットワークも，出力層にはソフトマックス関数を使った．学習に関しては，検証データを使って評価をした．ニューラルネットワークや深層学習では，モデルの過学習を防ぐ目的で，early stopping と呼ばれる方法をとることがある [1]．これは，訓練データと別に検証データを使って学習状態を監視し，分類誤差が上昇した時点で学習を打ち止めるものである．本実験では，この early stopping の基準を後述する AUC とした．すなわち，検証データの AUC が最も大きくなった点のパラメータをテストデータに用いた．

6.3 評価方法

実験の評価は適合率，再現率，適合率-再現率曲線，AUC により行った．再現率，適合率はそれぞれ，
 true-positive: 正しく正と予測した数，
 true-negative: 正しく負と予測した数，
 false-positive: 間違っ正と予測した数，
 を使って以下のように表せる．

$$\text{適合率} = \frac{\text{true_positive}}{\text{true_positive} + \text{false_positive}}$$

$$\text{再現率} = \frac{\text{true_positive}}{\text{true_positive} + \text{false_negative}}$$

また，再現率，適合率だけではなく，再現率-適合率曲線による評価も行った．各分類器が出力する連続値に対して閾値を変化させ，その閾値ごとの適合率，再現率をプロットした．

7 実験結果

7.1 適合率，及び再現率

適合率及び再現率を示す．

Abbreviated Name	SVM		NeuralNetwork		DeepLearning	
	P	R	P	R	P	R
Gupta_A	0.9715	0.2783	0.913545	0.32314	0.835141	0.392457
Sato_M	1	0.404	0.904192	0.5	0.77551	0.503311
Smith_J	0.8394	0.1756	0.818182	0.718626	Nan	0
Tanaka_K	1	0.461	1	0.965812	1	0.940171

Fig. 2 適合率, 再現率

適合率及び再現率は，分類器の分類結果に基づく．つまり，SVM では閾値を 0，ニューラルネットワーク，深層学習では閾値を 0.5 としている．深層学習の Smith_J において適合率が Nan, 再現率が 0 となっているのは，正しく正としたものも，間違っ正としたもの 0 であったためである．適合率において，深層学習は総じて悪い結

果となった．特に，Smith_J では true_positive が 0 件のため，適合率が計算できなくなっている．一方で，再現率に関しては，比較的高い値を保っていることがわかる．また，このときの各混合行列も以下に示す．混合行列は各列が予測値，各行が実際の真偽値を表す．

Abbreviated Name	SVM	Neural Network	Deep Learning
Gupta.A	0.67843	0.65704	0.59749
Sato.M	0.85705	0.7625	0.68017
Smith.J	0.63485	0.44332	0.39264
Tanaka.K	0.90315	0.919	0.86784

Table 2 AUC

7.2 適合率-再現率曲線, AUC

以下に，Gupta_A という著者名の適合率-再現率曲線を示す．また，AUC は，この曲線下の面積であり，理想的な分類器であれば，この値は 1 となる．

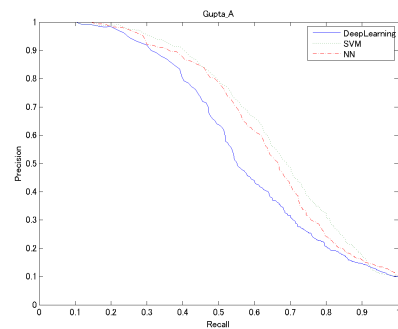


Fig. 3 Gupta_A

適合率，再現率曲線は，深層学習は全てのデータの全域にわたって悪い結果となった．AUC によって，そのことが確認できる．ニューラルネットワークが最も良い結果を示した例もあったが，次いで先行研究が高く，やはり深層学習は低い．

8 考察

SVM と比較して深層学習で性能が低い理由を考察する．

まず，データについて考察を加えるために，各訓練データのスパース性に着目する．

Abbreviated Name	素性数	平均素性数
Gupta.A	3444	19.7
Sato.M	1724	22.2
Smith.J	3652	17.7
Tanaka.K	1832	20.3

Table 3 各訓練データの素性数と，ペアインスタンスに含まれる平均素性数

本実験に使用したデータは、表 8 に示すように高次元でスパースな点が特徴的である。このように、1つのペアインスタンス中に存在する素性の数は非常に少なく、ほとんどの素性は 0 である。したがってこのようなデータは、ノイズの影響を非常に受けやすいデータだといえる。そのため、事前学習や、その後の学習がうまく行われない場合、容易に元の情報が失われてしまう。

もう一つの理由として考えられるのは、過学習の問題である。深層学習は、ネットワークが多層なため、モデルのパラメータ数が非常に多くなる傾向にある。したがって、訓練誤差が小さくなくても、汎化性能が落ちてしまう過学習の問題が起こりやすい。本研究では、この問題への対策として、AUC による early stopping を行ったが、この方法が有効でないことがわかる。

9 まとめ

5章の実験結果より、深層学習では、適合率、再現率-適合率曲線、AUC で SVM より悪い結果となっている。再現率では、何件か良い結果となっているものの、計算コストとの兼ね合いでニューラルネットワークのほうが実用面で有利である。したがって、今回の設定のもとでのペアワイズ分類における深層学習は有効とはいえない。

9.1 課題

今回の設定で明らかになった問題点は、データのスパース性及び高次性、モデル選択を含む学習の困難さである。データのスパース性及び高次性を回避するには、学習に適した何らかのペアインスタンスを見つけることである。また、スパース行列を扱う方法としては疎行列演算のライブラリを使う方法がある。疎行列演算を基本としたネットワーク計算ができれば、ある程度高次元でスパースなベクトルを効率的に扱える可能性がある。しかし、バックプロパゲーションの過程では、勾配は密行列になることが一般的であり、根本的な解決ではない。

モデル選択に関しては、深層学習に一般的な問題となる。すなわち、SVM と異なり、深層学習を含むニューラルネットワーク型の学習器は最適解が得られることが保証されていない。加えて、膨大なハイパーパラメータを持つ深層学習のモデル設計に関しては、Bengio[1]をはじめ様々な提案がなされているが、未だ完全な指針は定まっていない。

9.2 展望

本研究の設定では深層学習は有効ではないことがわかった。一方、データベース等の更新されていくデータ集合に対し、問題に適した類似度を機械学習で自動的に獲得することの重要性は増していくことと考えられる。そうした場合、やはり、追加学習可能な学習器が望ましい。今後の展望としては、他のオンライン学習型アルゴリズムで、ペアワイズ分類に有効なものを検証したい。

参考文献

- [1] Yoshua Bengio. Practical recommendations for gradient-based training of deep architectures. *CoRR*, Vol. abs/1206.5533, , 2012.
- [2] Christopher M. Bishop. *Neural Networks for Pattern Recognition*. Oxford University Press, Inc., New York, NY, USA, 1995.
- [3] Thorsten Joachims. [svmlight](http://svmlight.joachims.org/). <http://svmlight.joachims.org/>
- [4] ビショップ CM. パターン認識と機械学習:. 第 3 巻. シュプリンガー・ジャパン, 2008
- [5] [Deeplearningtoolbox](https://github.com/skaae/DeepLearnToolbox). <https://github.com/skaae/DeepLearnToolbox>
- [6] [Dblp](http://www.informatik.uni-trier.de/~ley/db). <http://www.informatik.uni-trier.de/~ley/db>
- [7] Stanford University. [Ufldl tutorial](http://ufldl.stanford.edu/wiki/index.php/UFLDL_Tutorial). http://ufldl.stanford.edu/wiki/index.php/UFLDL_Tutorial
- [8] Pascal Vincent, Hugo Larochelle, Isabelle Lajoie, Yoshua Bengio, and Pierre-Antoine Manzagol. Stacked denoising autoencoders: Learning useful representations in a deep network with a local denoising criterion. *The Journal of Machine Learning Research*, Vol. 9999, pp. 3371–3408, 2010.
- [9] 岡谷孝之, 齋藤真樹. ディープラーニング. http://www.vision.is.tohoku.ac.jp/files/9313/6601/7876/CVIM_tutorial_deep_learning.pdf
- [10] ディープラーニングによるパターン認識. 情報処理, Vol. 54, No. 5, pp. 500–508, may 2013.
- [11] 異なる例からの素性の組合せを用いたペアワイズ分類器の学習. 人工知能学会論文誌, Vol. 20, pp. 105–116, nov 2005.