Multimodal Extreme Learning Machineによる Wikipedia記事のマルチモーダル検索

立間淳司[†, a]　　青野雅樹[†, b]

† 豊橋技術科学大学大学院 情報・知能工学系

a) *tatsuma@cs.tut.ac.jp*　b) *aono@tut.jp*

概要　順伝播型ニューラルネットワークの一つである Extreme Learning Machine (ELM) は，入力層と隠れ層間の重みをランダムに決定し，隠れ層と出力層間の重みのみ最適化する．シンプルなアルゴリズムであることから，高速にネットワークを学習できる．本研究では，ELM をマルチモーダルデータの解析のために拡張した，Multimodal ELM (M-ELM) を提案する．隠れ層で抽出された二種類の多変量データの特徴量を，共通の接続層にて合成する．M-ELM は，Wikipedia 記事のマルチモーダル検索実験において，従来手法よりも優れた検索性能を得た．

キーワード　Extreme Learning Machine，マルチモーダル検索，ニューラルネットワーク，情報検索

1　はじめに

近年，ソーシャルメディアの普及により，Web 上には多種多様なマルチモーダルデータが増え続けている．Flickr や Youtube といった共有サイトでは，コメントやタグなどのテキスト情報が付与された画像や動画が投稿されている．また，Wikipedia には，解説文と画像が掲載された記事が増え続けている．このような状況から，画像を検索質問として文書を，文書を検索質問として画像を検索するといった，マルチモーダル検索技術が必要とされている．

マルチモーダルデータの共通する特徴を抽出する方法として，多変量解析の分野では，正準相関分析 (Canonical Correlation Ananlysis, CCA) が知られている．CCA は，二種類の多変量データの相関が最大となるような部分空間を解析する．マルチモーダル検索においても，Rasiwasia ら [2] は，CCA により抽出した低次元特徴量を用いて，Wikipedia 記事の文書による画像の検索，画像による文書の検索を実現した．

一方，近年，ニューラルネットワークにより，データから抽象度の高い表現を得る方法が，再び注目を集めている．Extreme Learning Machine (ELM) [1] は，順伝播型ニューラルネットワークの一つであり，入力層と隠れ層間の重みをランダムに決定し，隠れ層と出力層間の重みのみ最小二乗法により最適化する．ELM は，シンプルなアルゴリズムであることから，GPU など特別なハードウェアを必要とせず，高速にネットワークを学習できる．また，ELM には，様々な拡張アルゴリズムが提案されており，隠れ層を多層化した Hierarchical ELM (H-ELM) [4] などがある．

本研究では，ELM を，複数の多変量データを解析できるよう拡張した Multimodal ELM (M-ELM) を提案する．M-ELM は，Wikipedia 記事を対象としたマルチモーダル検索実験において，CCA に基づく手法よりも，優れた検索性能を得ることができた．

2　Multimodal ELM

2.1　アルゴリズム

図 1 に M-ELM のネットワーク構造を示す．H-ELM [4] の隠れ層と出力層の間に，共通の特徴を抽出するための接続層を加えた構造となる．

クラス分けされた二種類の多変量データの組 (X, Y) が与えられたとする．M-ELM では，入力層から隠れ層，接続層，出力層へと順に学習ステップを進む．

まず，入力層と隠れ層間および隠れ層間の重みを，H-ELM と同様に，自動符号化器の枠組みから学習する．入力層・中間層・出力層の 3 層のネットワークを考え，ランダムに決定した重み $\mathbf{a}_i$ とバイアス b_i，活性化関数 $g(\cdot)$ から，中間層の出力を得る．

$$h_i = g(\mathbf{a}_i^\top \mathbf{x} + b_i), \quad i = 1, \ldots, L$$

ここで，L は中間層のユニット数である．そして，出力層の出力が入力データとなるべく一致するように，中間層と出力層間の重み β を学習する．これは次式から求められる．

$$\beta = \left(\frac{I}{C} + H^\top H \right)^{-1} H^\top X$$

ここで，C は正則化パラメータであり，H は中間層の出力からなる行列である．出力層を取り除き，得られた重み β を，改めて入力層と隠れ層間および隠れ層間の重みとする．以上より，M-ELM における i 番目の隠れ層の

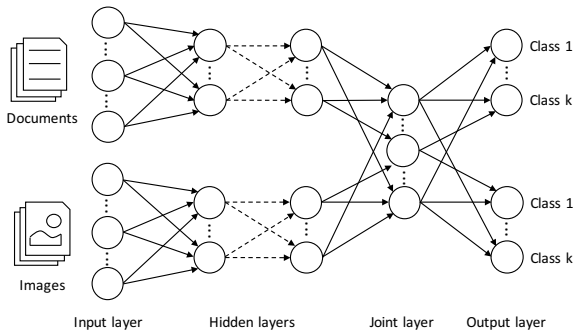

図 1 Multimodal Extreme Learning Machine の構造

出力 H_i は，重み β_i と $i-1$ 番目の隠れ層の出力 H_{i-1} から

$$H_i = g(\beta_i^\top H_{i-1})$$

で得られる．ここで，H_0 は入力層を示す．この処理を X と Y の双方で行い，それぞれの隠れ層の出力を得る．

次に，隠れ層と接続層間の重みの学習は，双方の隠れ層の出力が重み W_X，W_Y によって，接続層において共通の値を出力するように学習する．これは以下の最適化問題を解くことで得られる．

$$\underset{W_X,W_Y}{\operatorname{argmin}} ||H_X W_X - H_Y W_Y||^2$$

この問題は，接続層の出力の大きさに制約を加えることで，CCA と同様となる．

最後に，接続層と出力層間の重みを，ELM と同様に最小二乗法により学習する．

2.2 マルチモーダル検索への応用

Rasiwasia ら [2] は，CCA とロジスティック回帰によるマルチモーダル検索手法 Semantic correlation matching (SCM) を提案した．SCM では，まず，二種類の多変量データの組から，CCA を用いて共通の低次元表現を抽出する．次に，抽出した低次元表現を用いて，ロジスティック回帰により各クラスに対する事後確率を求める．そして，得られた事後確率を特徴量として，マルチモーダル検索を行う．M-ELM でもこれにならい，各クラスに対する帰属度を表す出力層の値を，特徴量として検索を行う．ここで，出力層において負の値となったものは，不要な特徴であると考えて 0 に打ち切る．

3 実験

データセットには，Wikipedia articles dataset (WAD) [2] を用いる．WAD には，10 種のクラスに分類された，文書と画像が組になった 2,866 記事が含まれており，2,173 記事が訓練データ，693 記事がテストデータとして提供されている．また，文書については Latent Dirichlet Allocation モデルによる 10 次元の特徴量が，画像については Bag-of-Visual-Words ヒストグラムによる 128 次元の特徴量が提供されている．従来手法との比較のため，本研究でもこれら特徴量を用いた．

表 1 MAP による検索性能の比較

Method	Image Query	Text Query	Average
CM	0.249	0.196	0.223
SM	0.225	0.223	0.224
SCM	0.277	0.226	0.252
M-ELM	**0.307**	**0.234**	**0.270**

M-ELM のネットワーク構成は，文書・画像ともに隠れ層を 3 層用意し，文書特徴量を入力とする隠れ層のユニット数はそれぞれ 40, 画像特徴量を入力とする隠れ層のユニット数はそれぞれ 512 とした．隠れ層の活性化関数には正弦関数を用いた．また，接続層のユニット数は 40 とした．

比較手法には，Rasiwasia らの SCM, Correlation matching (CM), Semantic matching (SM) を用いた．

表 1 は，Mean Average Precision (MAP) による各手法の検索性能をまとめたものである．M-ELM が，従来手法よりも優れた検索性能を得ていることがわかる．

4 おわりに

本研究では，二種類の多変量データの共通する特徴を解析する手法として Multimodal Extreme Learning Machine (M-ELM) を提案した．M-ELM は，Wikipedia 記事のマルチモーダル検索実験において，従来手法よりも優れた検索性能を得た．今後の課題は，接続層の学習法の改良，同様のネットワーク構造を持つ手法 [3] との比較があげられる．

謝辞

本研究の一部は，栢森情報科学振興財団および JSPS 科研費 26280038, 15K12027, 15K15992 の助成を受けたものです．

参考文献

[1] Huang, G. B., Zhu, Q. Y., Siew, C. K.: Extreme learning machine: theory and applications, Neurocomputing, 70 (1), pp. 489–501, 2006.

[2] Rasiwasia, N., Pereira, J. C., Coviello, E., Doyle, G., Lanckriet, G. R. G., Levy, R., Vasconcelos, N.: A new approach to cross-modal multimedia retrieval, Proc. of International Conference on Multimedia (MM'10), pp. 251–260, 2010.

[3] Srivastava, N. and Salakhutdinov, R.: Multimodal learning with deep boltzmann machines, J. Mach. Learn. Res., 15, pp. 2949–2980, 2014.

[4] Tang, J., Deng, C., and Huang, G. B.: Extreme learning machine for multilayer perceptron, IEEE Trans. Neural Netw. Learn. Syst., PP (99), 2015.