N-013

敦煌文書における機械学習を用いた字形分析

-既存分類手法との比較を中心として-

高見美友 *　藤田和弘*　三谷真澄*　入澤　崇*
Miyu TAKAMI　Kazuhiro FUJITA　Mazumi MITANI　Takashi IRISAWA

1　はじめに

従来から，敦煌文書の漢字に関して，藤枝[1, 2]は，写本の関連情報の整理・研究により得た成果から，文書の制作年代を分類する手法を提唱し，その手法は現在も使われている．この分類手法は，計測可能な情報と計測困難な情報の両方を，文書分類の基準として用いている．その計測困難な情報とは，紙質の手触りや字形に関する印象などであり，暗黙知として判断に用いられている．その暗黙知の字形に関する印象を定量化することを目的として，本研究では機械学習手法を用いた字形分析を行った．本研究では，敦煌文書の写本にある「不」の漢字画像に対して，主成分分析 (PCA, Pricipal Component Analysis)を適用し，各漢字画像のPCA展開係数を特徴量として用いた．そのPCA展開係数に対して，クラスタ分析としてGMM (Gausian Mixture Model)を用い，各クラスタと敦煌文書の写本との対応，藤枝が提唱した既存の分類手法と各クラスタとの対応について，考察を行った．なお，クラスタ数の決定には，BIC (Bayesian Information Criterion)を用いた．また，各クラスタの字形特徴を捉えるために，各クラスタの重心のPCA展開係数から再構成した画像を用いた．

2　データセット

本研究において作成した「不」の漢字画像のデータセット(15 クラス) を，表1に示す．「不」の漢字画像の選択基準は，HNG (Hanzi Normative Glyphs, 漢字字体規範史データベース)，IDP (International Dunhuang Project)，フランス国立図書館運営電子図書館のいずれかのデータベースに画像が格納されていることとした．

3　主成分分析(PCA)による特徴量

表1の画像に対して，主成分分析(Principle Component Analysis)を行い，PCA展開係数を求め，以後，それを特徴量として用いた．

表1　用いた敦煌文書のデータセット

クラス番号	文書名	書写年代	画像数
1	大樓炭經卷三(P2413)	589	50
2	賢劫經卷二(正倉院本)	610	57
3	妙法蓮華經卷五(P2334)	617	33
4	妙法蓮華經卷八(S2577)	7C末	35
5	妙法蓮華經卷四(S2157)	691	2
6	摩訶摩耶經卷上(P2160)	586	54
7	妙法蓮華經卷五(今西本)	671	44
8	妙法蓮華經卷三(守屋本)	675	30
9	妙法蓮華經卷六(P2195)	675	27
10	妙法蓮華經卷四(S312)	673	34
11	妙法蓮華經卷四(S3079)	671	60
12	妙法蓮華經卷四(S4551)	672	55
13	華嚴經卷十六(S2067)	514	49
14	華厳経巻35(P2110)	513	104
15	華厳経巻39(S9141)	513	2

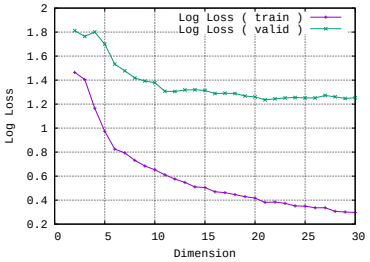

図1　PCA次元数と識別クロスエントロピー

まず，主成分分析の次元数を決定するために，次元数を変化させながら，主成分分析を行い，そのPCA展開係数を学習用データと検証用データに2分割し，ガウシアン・ナイーブベイズ識別器(GaussianNB)を学習用データで学習し，検証用データの識別実験を行い，分類クロスエントロピーを求めた．その結果を，図1に示す．図1より，次元数を22と決定した．また，その際の固有画像の一部を，図2に示す．図2を見ると，固有漢字画像は，2画目と4画目の位置や長さを反映していると判断できる．

* 龍谷大学，Ryukoku University

(a)第1成分　(b)第2成分　(c)第3成分　(d)第4成分

図2　固有画像

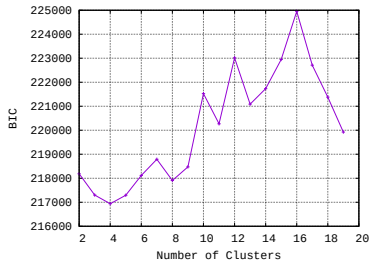

図3　クラスタ数とBIC

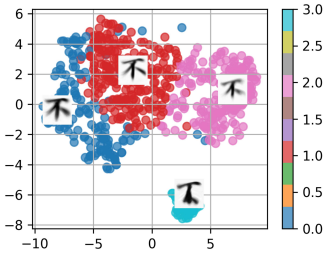

図4　クラスタ分析結果

4　GMMによるクラスタリング

PCA展開係数に対してGMM（ Gaussian Mixture Model ）を適用して，データセットを適切なクラスタに分割を行う．具体的には，クラスタ数を変化させながら，GMMによるクラスタリングを行い，クラスタリングの評価値としてBIC（ Bayesian Information Criterion ）を求めた結果を，図3に示す．図3より，BICが最小となるクラスタ数4を，これ以降の分析で用いる．

PCA展開係数をt-SNEにより2次元に次元削減を行った結果の散布図を，図4に示す．また，図4には，各クラスタの重心のPCA展開係数からPCA再構成を行った画像も示している．図4より，クラスタ0および1は隷書体の字形，クラスタ2は楷書体の字形，クラスタ3は特殊な字形であると判断した．また，クラスタ3は単一の写本から得た画像だけで構成されていた．

5　藤枝分期法との比較

藤枝分類A,B,Cと上記のクラスタ0,1,2との関係について，考察を行った．藤枝分類では，年代が下る順番に

表2　藤田枝分岐法との比較

	クラスタ0 (隷書相当)	クラスタ1 (隷書相当)	クラスタ2 (楷書相当)	合計
藤枝A	54（33%）	101 (63%)	5 (3%)	160
藤枝B	51（36%）	52 (37%)	37 (26%)	140
藤枝C	47（16%）	45 (16%)	195 (68%)	287
合計	152	198	237	587

A,B,Cとなっている．隷書や楷書等の書体は,年代ごとに特徴があるとされており,時代が下るにつれて,隷書から楷書へ変化する．藤枝分類では概ね，Aが隷書期に相当し，変遷期に相当するBを経て，楷書の確立後をCとしている．表1の文書のうち，藤枝分類が既知のものについて，本研究でのクラスタとの対応を，表2にまとめた．表2より，藤枝分類Aは隷書体に相当するクラスタ0と1の比率が高く，藤枝分類Bは字形の変遷期に相当するためか，クラスタの比率が0,1,2に分散した．藤枝分類Cでは楷書体に相当するクラスタ2の比率が高いことがわかる．時代により各クラスタの比率が変化しており，各書体毎に主たるクラスタが分かれた．クラスタの比率の変化は，藤枝分類や書体の変遷と一致しており，暗黙知として捉えられていた字形や書体には，定量的な違いがあると判明した．

6　おわりに

従来手法である藤枝分類法により行われていた敦煌文書の分類に対して，漢字の字形について主成分分析（ PCA ）およびGaussian Mixture Modelによるクラスタ分析を用いた機械学習手法による分類手法を提案し，両手法による分類結果の比較・検討を行った．本提案手法では，漢字の字形について，４つのクラスタに分類され，それぞれのクラスタが，隷書相当の字形，楷書相当の字形，特殊な字形に相当することがわかった．また，藤枝分類A,B,Cそれぞれとの対応関係についても示した．

今後は，「不」以外の漢字についても分析を行い，さらに敦煌文書の漢字の字形分析を進める予定である．

参考文献

[1] 藤枝晃:“ 中国北朝写本の三分期 「古筆と国文学」 ”, 八木書店（ 1987 ）

[2] 藤枝晃:“ 高昌殘影: 出口常順藏 トルファン出土佛典斷片圖録 ”, 法藏館（ 2005 ）