「マルチメディア，分散，協調とモバイル
(DICOMO2021) シンポジウム」 令和3年6月-7月

機械学習を用いた診療録の様式分類における アルゴリズムの一検討

小野 悟[1]　大石和佳[2]　Eric J. Grant[3]

概要：放射線影響研究所では，原爆被爆者およびその二世の方々の多大な協力を賜りながら，過去 70 年以上に亘り疫学的，分子生物学的研究を継続している．その過程において，これら研究協力者の臨床的情報を体系的な記録として保管するために診療録が作成されている．診療録を構成する情報の一部は紙媒体のみでしか保存されていないため，経年劣化による情報の滅失が懸念されており，構成する各様式等の光学スキャン方式を用いた電子化を検討している．しかしながら，これらの方式では読み取られた情報は画像情報として保存されるため，適切な分類や検索のためのメタデータの付加が望ましい．そこで，パイロットスタディとして保有する診療録の一部を光学スキャン方式によって画像化し，それらの様式に対して機械学習を用いた分類を試みた．本稿ではこの分類のために，決定木・K 近傍・SVM (Support Vector Machine) の 3 つのアルゴリズムを検証した．検証の結果，研究所が保有する診療録の画像分類においては，SVM が正解率等において優位な結果を示した．

Consideration of algorithms in the classification of medical chart using machine learning

Satoru Ono[1]　Waka Ohishi[2]　Eric J. Grant[3]

1. はじめに

1945 年 8 月に広島及び長崎に投下された原子爆弾によって多くの貴重な人命が失われ，戦後 75 年以上経過した現在でもその後遺症などに苦しむ人々は多い．広島および長崎の両県に拠点を持つ放射線影響研究所では，終戦後間もない 1947 年 3 月に開設された ABCC (AtomicBomb Casualty Commission) が行ってきた原爆放射線被爆者における放射線の医学的・生物学的影響の長期的調査を継承し，約 3.8 万人の直接被爆者および被爆者二世に対する臨床・疫学的，分子生物学的研究を継続している．

研究協力者である被爆者らは，定期的に研究所に来訪され，身体検査，血液検査，生理検査等所定の検査を受検されると同時に医師の問診を受ける．これらは一般的な病院での診察と同様に実施され，結果は診療録として保管される．診療録は，数値化可能な検体検査結果値等一部の情報が電子情報として併存されるが，全構成情報は体系的に編綴された紙媒体として管理されている．

現在，広島研究所には概数として直接被爆者の診療録が 20,074 冊，被爆者二世の診療録は 9,498 冊が保管されている．紙媒体で保管される診療録の中には 1940 年代に作成されたものもあり，経年劣化による記録内容滅失の懸念に加え，広島大学原爆放射線医科学研究所との共同研究に基づく原爆被爆医療や人文科学系の研究に寄与可能なデジタルアーカイブスの構築に関するコンソーシアム [1] のためのデジタル化が検討されており，紙媒体からの効果的な電子化手法について議論が進められている．

デジタル化の１手法として光学スキャン方式を検討した．しかしながら，全診療録を構成する各様式の編綴をすべて展開したうえで光学スキャンを行うためには多くの時間的

[1] 放射線影響研究所 情報技術部
Information Technology Depertment, Radiation Effects Research Foundation
[2] 放射線影響研究所 臨床研究部
Crinical Studies Depertment, Radiation Effects Research Foundation
[3] 放射線影響研究所 主席研究員
Associate Chief of Research, Radiation Effects Research Foundation

金銭的コストを要することから，光学スキャン方式を用いたデジタルアーカイブの有効性を確認するために直接被爆者および二世のカルテから任意に一部の診療録をその対象として抜粋して評価するパイロットスタディを行うこととした．光学スキャン対象として，直接被爆者診療録 24 冊，二世診療録 10 冊を任意に抽出した．構成する各様式の総数はそれぞれ約 7,005 枚と 1,256 枚である．すなわち，これらと同数の画像ファイルが光学スキャン方式によって生成されている．これらの画像ファイルへの意味づけの一環として診療録を構成する要素であるそれらの様式毎に機械学習を用いた分類を試みた．画像ファイルとして保存される多様の様式に対して適切な分類が可能であれば，様式毎に画像ファイルを分類保管したうえで適切なインデクシングが可能であると考えられる．本稿ではこの分類のために用いるアルゴリズムとして決定木・K 近傍・SVM(Support Vector Machine) の 3 つの手法について検証する．

本稿の構成は以下のとおりである．2 節では，研究所が保有するカルテの主要構成要素と今回分類対象とした様式について述べる．さらにそれらの画像ファイルの量子化方法と事前評価方法について考察しながら，評価を行うための環境について説明する．続く 3 節では，各アルゴリズムの評価結果について述べる．4 節でこれらの考察を行い，最後の 5 節をまとめとする．

表 1 カルテの主要構成シート

No.	シート名	手書	データ	対象
1	記録管理表 Record Control Sheet(RCS)	○		○
2	診断の総括 Diagnostic Summary(DS)	○		○
3	経過記録	○	○	
4	罹病及び入院に関する病気 Medical History(MH)	○		○
5	現症歴及び既往歴 System Review(REVIEW)	○		○
6	全身検査 Physical Examination(PHYE)	○		○
7	検査データ（血液、尿、便）		○	
8	放射線科検査	○	○	
9	心電図検査所見 ECG Report(ECGR)	○		○
10	心電図検査結果シート（波形）	○		
11	同意書		○	
12	薬剤調査票		○	
13	骨密度検査		○	
14	健診結果報告書		○	
15	各種問診票	○	○	

2. 手法

2.1 対象様式の選択

研究所の診療録は 15 種類の主要な様式から構成されている．この様式の中から，今回は表 1 の区分に従って分類対象となるものを選択した．また，選択対象からは様式に記載された情報が何らかの形で電子化されているものは除外した．今回選択した様式中には，同じ様式でも歴史的な変遷を経てフォーマットに大小の変化があるものや，各様式間に類似性があるものが存在する．すなわち，同じ目的の様式でも古いものと新しいものでは，フォーマットが全く異なるものや，別の様式でも様式的には表形式のような非常に似通ったものが存在していることから，画像認識として各様式を一意に判別することは困難であることが予想された．なお，今回対象とした 6 種類のシート中，1. 記録管理表のみシート色がピンク色であり，他のシートはすべて白色となっており，画像ファイルの構成要素としての画素値が他の様式と比較した場合，大きく異なるものであることに留意されたい．心電図検査所見に添付されている波形情報は電子化されていないため，波形チャートは電子化対象となるが今回は除外している．これらは将来的に DICOM[3] 化することを想定しているためである．

2.2 画像ファイルの量子化

光学スキャン方式で電子化された画像ファイルは JPEG 形式である．ファイル毎に画素サイズが異なっていたため，まず全ファイルの画素サイズを同一長に成形した．原本の用紙サイズは B5 版であるため，大きく画素サイズが異なっていることはなく，余白のトリミング程度の処理によって同一化することができたことから，オリジナルの画像情報は損なわれていない．画素サイズは 3200 × 4210 とした．RGB ビット深度は 24bit である．この成形に基づき，3 原色の画素値と各ラスタ情報から構成される二次元配列情報を一次元配列として変換した後，これらの成分を固定要素長の CSV ファイルとして保存する．例えば，ある一つのシートの画像ファイルを CSV ファイルに変換した場合，概ね 150MByte 程度のテキストファイルが生成されることになる．また，CSV ファイルに保存された 3 原色の画素値の正当性を検証するために，PPM 形式の画像ファイルを同時に生成し，オリジナルの JPEG 画像との相違を目視で直感的に確認できるようにした．これらの手法によって，生成された CSV ファイルが有する成分値のおおまかな正当性を短時間で検証することができる．PPM 形式の画像ファイルは，プレーンテキストとしてヘッダ情報を持ち，各画素値そのものはバイナリで格納するフォーマットとなっている．今回はマジックナンバーとして P6，フルカラーバイナリモードでの生成とした．

2.3 主成分分析による事前評価

6 種の様式分類に関する事前評価のために各様式に対して主成分分析を試みた．2 要素で評価した結果を図 1 に示す．○で囲まれた記録管理表 (RCS) に関しては，用紙色が他の様式と異なるピンク色であるため，2 次元の主成分分析でもはっきり分類可能であることが示唆されている．しかしながら，他の様式に関しては，2 次元上の評価では分類が困難であることが想定されたため，図 2 に示す寄与率を求めた．図 2 より概ね 80 程度の主成分を用いることで 8 割前後の寄与率が得られることがわかった．よって，少なくとも 8 割前後の分類正答率が得られると仮定した．

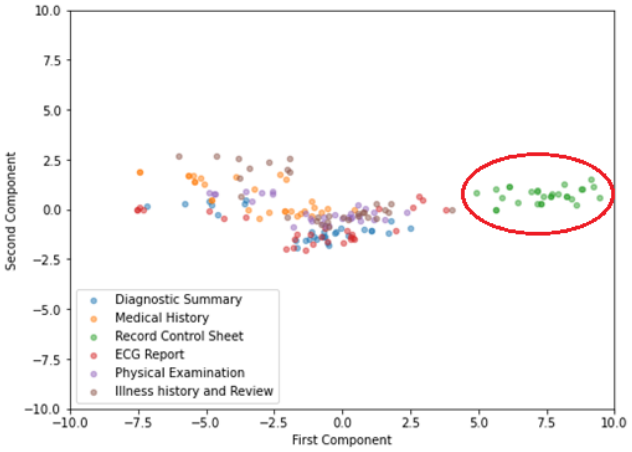

図 1 主成分分析による評価

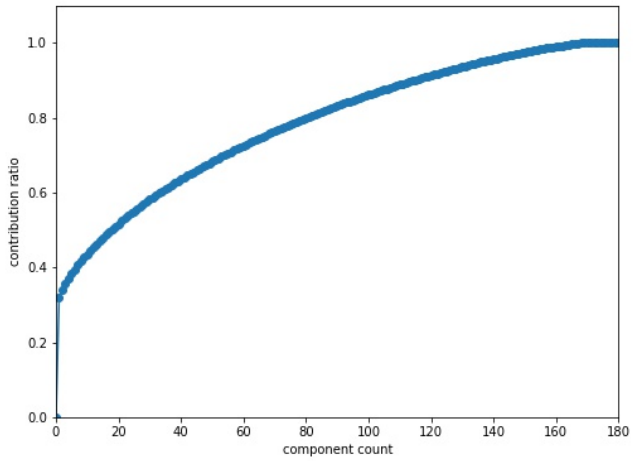

図 2 寄与率

2.4 評価環境

評価環境として，Dell 社製 PowerEdge R840 (Intel Xeon Gold 6126@2.6GHz × 48CPU, 1.5TB Memory) の物理マシンを用いた．この物理マシンに vSphere ESXi-6.7.0 の仮想マシンとして UBUNTU 20.04.2 LTS をインストールし，Python で利用することができる機械学習のオープンソースライブラリである scikit-learn を利用した．scikit-learn で扱うことが可能なモデルから適切なアルゴリズムを選択することができるチートシート [2] を参考にしたうえで，今回の分類では決定木・K 近傍・SVM の 3 つのアルゴリズムを評価することとした．

教師データとして各様式からそれぞれ 50 シート（合計 300 シート）をランダムに抽出した．また，機械学習の分類能力の評価のために 6 種類のいずれの様式にも該当しない 50 シートをランダムに選択し，教師データに加えた．すなわち画像ファイルとして 350 ファイルを教師データとする．検証データとして 6 様式から 18 シートをランダムに選択した．それぞれの比率は概ね 75:25 として設定している．これらのファイルを 2 節で述べた通りの手順で CSV ファイル化した．教師データの CSV ファイルは約 40GByte となった．一方，検証データの CSV ファイルは約 12GByte となった．

3. 各アルゴリズムの評価結果

3.1 決定木

評価結果として混同行列を表 2 に示す．正解率，適合率，再現率，F 値を表 3 に示す．表 2 において，縦軸が真のラベル，横軸が予測ラベルである．対角線上のセルは正しく分類された観測値に対応する．以後，K 近傍・SVM に示す表も同様とした．

表 2 混同行列

	RCS	DS	MH	OTH	ECGR	PHYE	REV
RCS	15	0	0	3	0	0	0
DS	1	11	4	1	0	1	0
MH	0	4	9	3	0	0	2
OTH	0	0	0	0	0	0	0
ECGR	0	2	0	2	11	0	3
PHYE	0	0	3	2	0	12	1
REV	0	1	4	3	0	1	9

表 3 正解率

	Precision	Recall	F1-score
RCS	0.94	0.83	0.88
DS	0.61	0.61	0.61
MH	0.45	0.5	0.47
ECGR	1	0.61	0.76
PHYE	0.86	0.67	0.75
REVIEW	0.6	0.5	0.55
Accuracy	0.62		

3.2 K 近傍

評価結果として混同行列を表 4 に示す．正解率，適合率，再現率，F 値を表 5 に示す．

表 4　混同行列

	RCS	DS	MH	OTH	ECGR	PHYE	REV
RCS	18	0	0	0	0	0	0
DS	1	11	5	0	1	0	0
MH	0	0	18	0	0	0	0
OTH	0	0	0	0	0	0	0
ECGR	0	0	2	4	12	0	0
PHYE	0	0	6	4	1	7	0
REV	0	0	9	1	1	0	7

表 5　正解率

	Precision	Recall	F1-score
RCS	0.95	1	0.97
DS	1	0.61	0.76
MH	0.45	1	0.62
ECGR	0.8	0.67	0.73
PHYE	1	0.39	0.56
REVIEW	1	0.39	0.56
Accuracy	0.67		

3.3　SVM

評価結果として混同行列を表 6 に示す．正解率，適合率，再現率，F 値を表 7 に示す．DS（診断の総括）を 1 件だけ OTH（その他）に誤分類したのみで，他は全て完璧な分類を行うことができた．

表 6　混同行列

	RCS	DS	MH	OTH	ECGR	PHYE	REV
RCS	18	0	0	0	0	0	0
DS	0	17	0	1	0	0	0
MH	0	0	18	0	0	0	0
OTH	0	0	0	0	0	0	0
ECGR	0	0	0	0	18	0	0
PHYE	0	0	0	0	0	18	0
REV	0	0	0	0	0	0	18

表 7　正解率

	Precision	Recall	F1-score
RCS	1	1	1
DS	1	0.94	0.97
MH	1	1	1
ECGR	1	1	1
PHYE	1	1	1
REVIEW	1	1	1
Accuracy	0.99		

4.　考察

図 3 に 3 つのアルゴリズムの正解率を比較した．機械学習を用いた研究所の診療録の様式分類においては，SVM が有意に機能することがわかった．検証のために必要となる計算機リソースとして，当初 256Gbyte の主記憶容量を有する物理マシンを用いたが，全く容量が足りず，上限である 1.5Tbyte まで主記憶容量を拡張したうえで本評価を完遂したことを付け加えておく．また，計算機リソースの低減のために配列として展開された画素情報から平均，分散，偏差などの特徴点を利用することは可能であるが，正解率は著しく低下し，実質的には評価できないものであることもわかっている．これらについては，別の機会に議論したい．

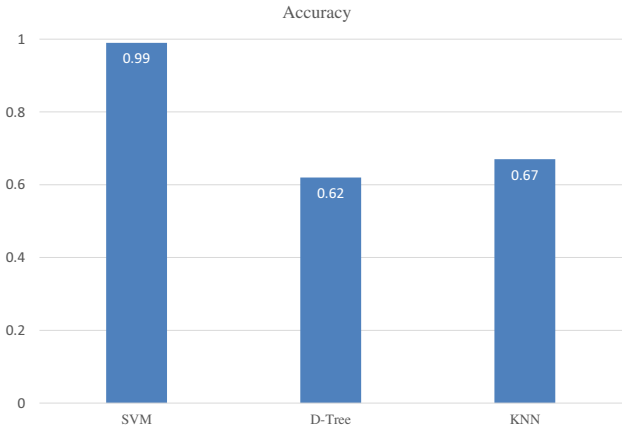

図 3　正解率の比較

5.　おわりに

今回の検証では，SVM において 99%という高い正解率が導出されたが，研究所が保有するカルテの全冊は約 900 万枚のシートから構成されており，仮に実運用時において 99%の正解率が得られたとしても，概算では約 9 万枚の誤分類が想定されることになる．これらは目視で分類されなければならない．実運用のためには 100%の正解率を目指すことが望ましいと考えており，今後は深層学習の活用を検討する．また，同時に実践的な正解率を維持したうえで，計算機リソースを低減するための手法についても検討を進めたい．

参考文献

[1] 中国新聞デジタル, 【ヒロシマの空白 被爆 75 年】放影研と原医研、被爆関連資料を共同活用へ デジタル化、閲覧可能に, コラム・連載・特集, 2020/12/30.
[2] scikit-learn algorithm cheat-sheet, Available on: `https://scikit-learn.org/` （参照 2021-02）.
[3] 木村通男, 医療情報の IT 化と医療情報学–電子カルテとどう付き合うか (8) 医療情報システムの相互運用性 (3) データ形式–HL7,HL7 CDA,DICOM で医療情報システムの標準化, 医学のあゆみ 222(2), pp147-154, 2007.7.14