ポスター | 知識工学／医療データ解析

ポスター1
知識工学／医療データ解析

2019年11月22日(金) 14:50 ～ 16:20 ポスター会場2 (国際展示場 展示ホール8)

[2-P2-1-05] 機械学習による退院時サマリからの DPC分類の推測

○木村 知広[1]、津本 周作[2] （1. 島根大学医学部医療サービス課, 2. 島根大学医学部医学科医療情報学）

キーワード：Text mining, Random Forest, Deep Learning, Discharge Summary

【背景】診療情報はビッグデータとして蓄積され、統計手法を用いて新たな知識を発見し，診療等に役立てることが可能となってきている。しかし，その多くは入力・保存や統計処理が容易な数値、画像データが対象で，カルテ記録などの診療文書は見過ごされてきた。

【目的】診療文書を対象とした分析法を確立することは，医療の質向上に大きく寄与すると考える。退院時サマリから疾患別の重要語を抽出し，これを元にサマリ文章から疾患名を特定できるか検討した。

【方法】キーワードとターゲットとの対応付けが弱く，分類精度がそれほど上がらないという，従来のテキストマイニングによる分類学習の問題に対して，本研究では，対応分析による DPC毎のキーワード選定を行い，機械学習で用いられる手法を適用，キーワードによる DPCコーディングを行う分類器を構築した。
マイニングのプロセスとして，抽出した退院時サマリに形態素解析を行い，キーワードの分割表を作成。次いで，対応分析を行い，各キーワードと DPCについて布置座標を与え， DPCとキーワード間のユークリッド距離を計算し， DPC毎に距離の値によってキーワードのランク付けを行う。それを用いて，文書内のキーワードの有無についての表形式のデータ集合を生成し，決定木などの分類学習手法を用いて分類器を構築する。その後、反復交差検証法により，分類の正答率を比較した。

【結果】検証結果から， SVM，決定木が最大78%程度の正答率であるのに対し，深層学習が86%以上， Random Forestでは90%以上の正答率を示した。

【考察】高い正答率を示しており，テキストマイニングによる診療文書内容からの疾患の特定・分類などの可能性が示唆された。この結果を踏まえて，退院時サマリの記載内容から DPCコードが推定できるとすれば，実際に退院患者に付与された DPCコードが正しいかチェックを行うなど応用が可能と考える。

©一般社団法人 日本医療情報学会

機械学習による退院時サマリからの DPC 分類の推測

木村　知広[*1], 津本　周作[*2]
*1 島根大学医学部医療サービス課, *2 島根大学医学部医学科医療情報学

Machine Learning based Prediction of DPC from Discharge Summaries

Tomohiro Kimura[*1] , Shusaku Tsumoto[*2]
*1 Division of Medical Service, Faculty of Medicine, Shimane University ,
*2 Department of Medical Informatics, Faculty of Medicine, Shimane University

This paper proposes a method for construction of classifiers for discharge summaries, composed of the following five steps First, morphological analysis is applied to a set of summaries and a term matrix is generated. Second, correspondence analysis is applied to the classification labels and the term matrix and generates two dimensional coordinates for all the terms and labels. Third, by measuring the distances between categories and the terms, ranking of key words is generated. Fourthly, keywords are selected as attributes according to the ranks, and training examples for classifiers will be generated. Finally, machine learning methods are applied to the training examples. Experimental validation shows that random forest achieved the best performance and the second best was the deep learners, but decision tree methods with many keywords performed only a little worse than neural network or deep learning methods.

Keywords: Text mining, Random Forest, Deep Learning, Discharge Summary

1. 背景および目的

電子カルテシステムの普及によって，診療情報は電子的データとして蓄積されつつある。島根大学医学部附属病院においても，オーダエントリシステムの導入から始まり，現在，診療情報のほぼすべての記録を保存・閲覧可能なシステムに発展した。一日あたり平均でおよそ 352.8MB のデータ(画像等除く)が日々蓄積されており，個々の患者の診療の記録であると同時に，ビッグデータとしての役割を担い，様々なデータ解析の手法を用いて新たな規則や知識を発見し，診療・研究・病院管理に役立てることが可能となってきている。

しかし，それらの研究の多くは医療情報の中でも，入力・保存や統計処理が容易な数値データや画像データが対象であり，医師のカルテ記録などの診療文書の解析に関する研究は少ない[1,2]。それは，従来の自然言語処理手法が診療文書の特殊性に対応できなかったことなどが理由として挙げられる。

医療文書の分析法を確立することは，医療プロセスの意思決定や予防医療の支援に期待されており，医療の質向上に大きく寄与するものと考える。その分析の一つとして電子化した退院時サマリからテキストマイニングによって疾患別の重要語を抽出し，これをもとに退院時サマリの文章から疾患名を特定できるかについて検討した。

2. 方法

診療文書には医療における用語の標準化がなされていないこともあり，多種多様な用語が混在し，解析しにくいことはよく知られているが，退院時サマリは各症例の診断・治療経過を時系列的に必要な情報を記載している。一般の診療記録に比べ，キーワードの冗長性が少なく，文自体も説明文の形式で省略は少なく，文章間の比較を含めて，横断的解析に適している。[3,4,5]

これまでテキストマイニングによる分類学習では，テキストからデータセットを作成し，それらを用いて，決定木・SVM 等の手法を用いる方法が提案されてきた[6]。ただし，従来の手法では，キーワードの選択は，キーワードの出現頻度のみに依存しているため，キーワードとターゲットとなる概念との対応付けが弱く，分類精度がそれほど上がらないことが指摘されている。本研究では，退院時サマリのテキストに形態素解析を適用後，分類に必要な変数選択を目的とした対応分析による DPC 毎のキーワードの選定を行い，選ばれたキーワードに機械学習で用いられる手法を適用し，キーワードによる DPC コーディングを行う分類器を構築した。分類器構築後，反復交差検証法を用いて，実際の退院時サマリを用いて，その分類の正答率を比較した。

プロセスは図 1 の通りである。

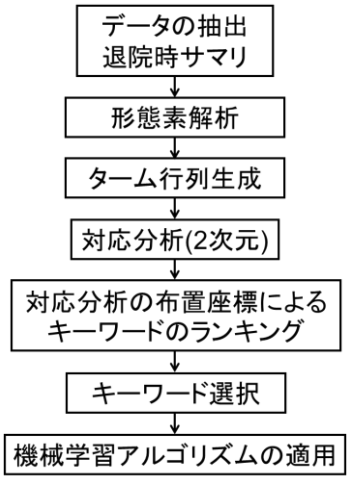

図 1　分類器構築プロセス

まず，電子カルテシステムから退院時サマリのテキストデータを抽出するプログラムを作成し、抽出。その後、テキストデータに対して，対象となる言語の文法や単語の品詞情報をもとに，文章を形態素(単語が意味を持つ最小の単位)に分解する形態素解析を行い，キーワードに関する分割表を作成する。次いで，対応分析(2 次元) を行って，各キーワードとDPCについて布置座標を与える。これらの布置座標について，DPC とキーワード間のユークリッド距離を計算し，DPC 毎に距離の値によって，キーワードのランク付けを行う。ランク付けされたキーワードを用いて，文書内のキーワードの有無についての表形式のデータ集合を生成する。生成されたデータを用いて，分類学習，ここでは決定木，SVM，深層学習などの手法を用いて，分類器を構築する。

形態素解析以降のプロセスは，図 2 の形の R-script で記述できる。

```
# 形態素解析 タート行列生成
library (RMeCab)

targetDir <- "csvdata"
test <- docMatrix(targetDir, pos = c("名詞","形容詞","助詞"), weight = "tf")
basedata <- t(test)
rownames(basedata) <- substr(rownames(basedata),1,14)


# 対応分析の布置座標によるキーワードのランキング
txt.list <- list.files(path = "csvdata", pattern = NULL, all.files = FALSE, full.names = FALSE, recursive = FALSE)
cnt <- length(txt.list)
data_t <- vector()
tbl.list <- vector()
colname <- c("Term")
flg <- 0
for (txt.file in txt.list){
      file <- paste("data\\",txt.file,sep="")

      ans <- RMeCabFreq(file)
      sortlist <- order(ans$Freq, decreasing=TRUE)
      ans1 <- ans[sortlist,]
      ans2 <- ans[ans$Freq > 10 ,]
      ans3 <- subset(ans2, (ans2$Info1 == "名詞" & ans2$Info2 != "数") | (ans2$Info1 == "形容詞"))

      file.name <- strsplit(txt.file, ".txt")[[1]]
      work <- ans3[,c(1,4)]

      work2 <- work
      rad <- work[duplicated(data.frame(work$Term)), ]$Term
      for (i in rad){
            wk <- subset(work, work$Term == i)
            wk2 <- c(i, sum(wk[,2]))
            work2 <- subset(work2, work2$Term != i)
            work2 <- rbind(work2, wk2)
      }

      if (flg == 0)(data_t <- work2) else (data_t <- merge(x=data_t, y=work2, by="Term", all=T))
      flg <- 1
      colname <- c(colname, file.name)
}
colnames(data_t) <- colname
data_t[is.na(data_t)] <- 0
DATA <- data_t[,2:(cnt+1)]
rownames(DATA) <- data_t[,1]
DATA <- data.matrix(DATA)

X <- 250    #キーワード数を指定
w.keyword.list <- vector()
cc <- DATAX.coa[[1]]
rscore <- DATAX.coa[[2]]
cscore <- DATAX.coa[[3]]
for (i in (1:nrow(cscore))){
      DATA <- sqrt(((rscore[,1]*cc[1]-cscore[i,1]*cc[1])^2)+((rscore[,2]*cc[2]-cscore[i,2]*cc[2])^2))
      sortlist <- order(DATA)
      w.keyword.list <- c(w.keyword.list, names(DATA[sortlist][1:X]))
      fn <- paste(rownames(cscore)[i],".txt",sep="")
      sink(fn)
      print(DATA[sortlist])
      sink()
}
keyword.list <- unique(w.keyword.list)
S <- vector()
for (j in keyword.list){
      S <- c(S, which(basedata[1,] == j))
}
data_t <- basedata[,c(1,S)]


# 機械学習アルゴリズムの適用
library(randomForest)

rowdata <- nrow(data_t)
rate <- vector()
for (I in 1:100){
      random_ids <- sample(1:2,rowdata,replace=TRUE)
      list <- 1:2
      result_predictx <- vector()
      predicting.setx <- vector()
      for (i in 1:2){
            training.set <- subset(data_t, random_ids %in% list[-i])
            predicting.set <- subset(data_t, random_ids %in% c(i))
            data_rp <- randomForest(Group~., data=training.set, na.action="na.omit")
            result_predict <- predict(data_rp, predicting.set, type="class")
            result_predictx <- c(result_predictx, as.vector(result_predict))
            predicting.setx <- c(predicting.setx, as.vector(predicting.set$Group))
      }
      #予測結果と正解との比較
      result <- table(result_predictx, predicting.setx)
      result
      #正答率
      rate <- c(rate, sum(diag(result))/sum(result))
}
```

図 2 分類器 R-script

3. 実験

島根大学医学部附属病院に 2015 年度に蓄積された退院時サマリのうち，DPC コードで上位 20 までの症例のものを抽出した(表 1)。データ抽出には，電子カルテシステムで採用されている InterSystems のデータベース Caché [8]を使用し，抽出プログラムを作成の後、抽出を行った。

表 1 2015 年度上位 20 症例

No	DPC 名称（DPC コード）	件数
1	白内障，水晶体の疾患 手術手術あり 重症度等片眼 (020110xx97xxx0)	445
2	白内障，水晶体の疾患 手術手術あり 重症度等両眼 (020110xx97xxx1)	152
3	2型糖尿病（糖尿病性ケトアシドーシスを除く。）(100070xxxxxxxx)	145
4	肺の悪性腫瘍 手術手術あり 処置等2なし (040040xx97x0xx)	131
5	子宮頸・体部の悪性腫瘍 手術手術なし 処置等24あり 副傷病名なし (12002xxx99x40x)	121
6	肺の悪性腫瘍 手術手術なし 処置等1あり 処置等2なし 副傷病名なし (040040xx99100x)	120
7	子宮の良性腫瘍 手術腹腔鏡下腟式子宮全摘術等 (120060xx02xxxx)	111
8	肺の悪性腫瘍 手術手術なし 処置等1なし 処置等24あり (040040xx9904xx)	110
9	妊娠期間短縮，低出産体重に関連する障害（出生時体重 2500g 以上） 手術手術なし 処置等2なし 副傷病名なし (140010x199x00x)	110
10	肘，膝の外傷（スポーツ障害等を含む。） 手術腱縫合術等 (160620xx01xxxx)	99
11	全身性臓器障害を伴う自己免疫性疾患 手術手術なし 処置等2なし (070560xx99x0xx)	96
12	非ホジキンリンパ腫 手術手術なし 処置等24あり 副傷病名なし (130030xx99x40x)	94
13	肺炎，急性気管支炎，急性細気管支炎（15 歳以上） 手術手術なし 処置等2なし (040080x099x0xx)	86
14	肺の悪性腫瘍 手術手術なし 処置等1なし 処置等2なし (040040xx9900xx	85
15	慢性腎炎症候群・慢性間質性腎炎・慢性腎不全 手術手術なし 処置等1なし 処置等2なし 副傷病名なし (110280xx99000x)	83
16	肝・肝内胆管の悪性腫瘍（続発性を含む。） 手術その他の手術あり 処置等2なし (060050xx97x0xx)	82
17	胆管（肝内外）結石，胆管炎 手術限局性腹腔膿瘍手術等 処置等2なし 副傷病名なし (060340xx03x00x)	82
18	脳梗塞（JCS10 未満） 手術手術なし 処置等1なし 処置等23あり 副傷病名なし (010060x099030x	80
19	網膜剥離 手術手術あり 重症度等片眼 (020160xx97xxx0)	75
20	胎児及び胎児付属物の異常 手術子宮全摘術等 (120180xx01xxxx)	73

抽出した退院時サマリのデータを解析ツール R のパッケージ RMeCab [9]を用いて，形態素解析を実施する。形態素解析を行った結果から単語の出現頻度を求め，退院時サマリ毎の集計したデータセットを作成する。

その後，作成したデータセットを元に，対応分析(R パッケージ MASS を使用: R3.4.4) を行い，各 DPC について，キーワードのランク付けを行った。対応分析結果の散布図を図 3 に示す。各 DPC 間で重複したキーワードについては重複を削除した。

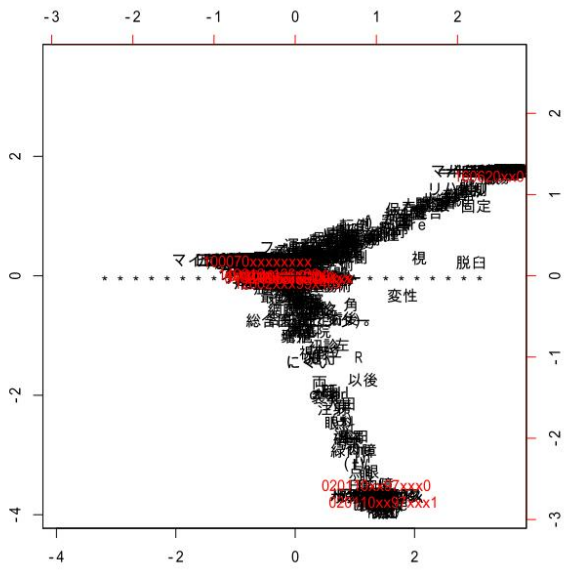

図 3　対応分析結果（散布図）

キーワードのランク付けにおける対応分析と TF-IDF の比較の一部を表 2 に示す。キーワードの選択が全く違うものとなっていることが分かる。

表 2　キーワードのランク付けにおける
対応分析と TF-IDF の比較（上位 4 症例）

	1. 白内障(片眼)		2. 白内障(両眼)		3. 2 型糖尿病		4. 肺悪性腫瘍	
順	対応分析	TF-IDF	対応分析	TF-IDF	対応分析	TF-IDF	対応分析	TF-IDF
1	in	IOL	眼	IOL	ケトン	mg	陰部	GIA
2	VA	PEA	(+),	PEA	CTR	ml	離床	Endo
3	下方	サンベタゾン	成熟	(Dr 名)	hr	時	バック	ics
4	DC	左	アバスチン	+	亢進	g	膣	葉
5	点眼	+	硝子	白内障	FDP	血糖	閉創	purple
6	硝子	右	Allergy	施行	雑音	性	グレード	Min
7	成熟	ジクロード	下方	trouble	跳痛	症	size	後
8	アバスチン	ベガモックス	In	視力	降圧	歴	growth	Lung
9	(+),	白内障	VA	退院	膵炎	入院	蒸留	肺
10	#.	月	The	左	閉経	病	キング	diag

分類器の構築には，以下の機械学習の方法を使用した。

・決定木
・SVM（サポートベクターマシン）
・RandomForest
・深層学習（Deep Learning）

分類器構築のプラットフォームとして，R3.4.4 を用い，決定木には rpart [10], SVM には kernlab [11], Random Forest には RandomForest [12]，深層学習については darch [13]を用い，darch のパラメータとしては，中間層(10) および中間層(10, 5)，中間層(20)，中間層(20, 10)，中間層(40, 20) ，反復学習回数を 100 とした。今回はどのパッケージについても，中間層を 2 層にした場合以外は，デフォルトのパラメータ設定を利用した。

次に，構築した分類器の性能評価については，データ集合をランダムに 2 分割し，片方を訓練標本，もう一方をテスト標本として正答率を算出することを 100 回繰り返し，平均正答率を算出した(repeated 2-fold cross validation[7])。

選択するキーワード数は 1 位のみから 1000 位までそれぞれについて，性能を評価した。

対応分析，分類器構築，性能評価については，Fujitsu PRIMERGY RX2530 M2 (Xeon プロセッサー E5-2690v4 (2.60GHz), 32GB メモリ) を用いた。R のバージョンは 3.4.4 を使用した。

4. 解析結果

各手法のキーワード数別による正答率のプロットを図 4 に示す。

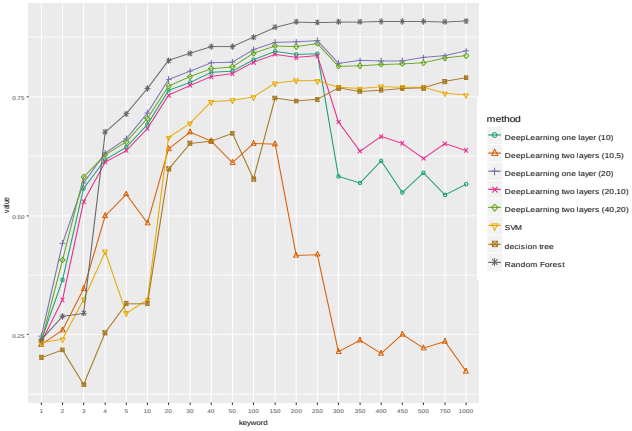

図 4　各手法のキーワード数別による性能評価結果

キーワード数の増加に伴い，どの手法も単調に増加しているが，キーワード数 150 付近でほぼ横ばいとなった。キーワード数 4 以上では，Random Forest の正答率が最も高く，他の手法よりも高い分類精度を示した。次いで，深層学習の精度が高いが，2 層(10,5) は 1 層に比べて性能が劣っていた。その他の決定木，SVM は，キーワード数の増加とともに，精度が上昇していくものの，Random Forest，深層学習ほどの精度の高さは出ていない。

5. 考察

Random Forest の正答率の高さが特筆される。いずれの手法もデフォルトのパラメータ設定を利用していることから、チューニング対象となるハイパーパラメータが少なく、デフォルトのパラメータで十分よく機能するとされる Random Forest の精度が高くなったと考えられる。キーワード数による正答率の変化については，SVM，深層学習ともにキーワード数 250 がピークで，SVM の正答率が 78.3%，深層学習が 86.8%となり，10%近い性能の差が得られた。これは，SVM がキーワードの線型結合によるのに対し，深層学習では，非線形結合が主となることに起因すると考えられ，キーワードの結合の性質，す

なわち線型とは言えない関係性がその精度を考える上で重要であることを暗示している。深層学習は，キーワード数 250 までは，中間層の違いがあっても，同様の分類精度の推移を示している。これは中間層ニューロン数の調整によって精度が高くなることを示している。この実験条件下では，中間層が増加した方が精度は上昇し，中間層のニューロン数を増加させることで，Random forest の正答率を越える可能性があるか検証を進める必要がある。

表 3 にその結果の一部を示した。正答率の高いキーワード数を 250 と固定し，同様の交差検証法で正答率を評価している。

表 3　中間層ニューロン数による正答率変化

キーワード数	中間層	平均正答率
250	1 層(10)	0.840
250	1 層(20)	0.868
250	2 層(10,5)	0.418
250	2 層(20,10)	0.836
250	2 層(40,20)	0.862
250	2 層(40,30)	0.860
250	2 層(80,20)	0.867
250	2 層(80,30)	0.871
250	3 層(80,40,20)	0.850
250	3 層(160,80,40)	0.870

以上のごとく，高次の関係性を想定することで性能が上がる可能性があるが，その性能向上は数パーセント程度にとどまっている。このあたり，中間層の設定に試行錯誤しているが，中間層を適切に調整すれば，Random forest と同等の性能あるいはそれ以上が期待できるのかもしれない。

実際に分類器がどのように誤判別しているかを表 4 に示す。これは深層学習の誤判別表であるが，白内障の片眼（020110xx97xxx0）と両眼（020110xx97xxx1）で誤りが多いことが際立っている。白内障と他の疾患の分類が大きく効いているため，片眼と両眼の区別がついていないことが考えられる。表 5 に白内障の実際の記載例とその判別結果をいくつか挙げる。

表 5　白内障の記載例

No	記載内容	正解	予測	判別
1	左白内障で H28 年 2 月 24 日入院。2 月 25 日 L)PEA+A-vit+IOL 施行（US 時に後嚢破損あり。IOL out 固定）。2 月 26 日退院し、以後外来受診予定である。ベガモックス、サンベタゾン、ジクロード左 4 点中である。	片眼	片眼	○
2	右白内障で H28 年 3 月 7 日入院。3 月 8 日 R)PEA+IOL 施行。3 月 9 日退院し、以後外来受診予定である。ベガモックス、サンベタゾン、ジクロード右 4 点中である。	片眼	片眼	○
3	○○医院より、両眼の霞で紹介。近医で白内障・緑内障を指摘された。 視力右（0.6）左（0.6）、眼圧右 14 左 14（GAT),右 13.0 左 11.7(NCT) H28.2/4 R)PEA+IOL, 2/9 R)PEA+IOL 施行。 術後経過良好で 2/10 退院となった	片眼	両眼	×
4	両白内障で H27 年 4 月 8 日入院。4 月 9 日 R)PEA+IOL 4 月 14 日 L)PEA+IOL を施行した。4 月 15 日退院し、以後は外来受診予定である。点眼はベガモックス、サンベタゾン、ジクロードを両 4 点中である。	両眼	両眼	○
5	□□市立より紹介。術前視力右(0.5)左(0.4)。 DM+、HT+、Stroke+ 【予定】11/18 入院、11/19 左白内障、11/24 右白内障、11/25 退院 【レンズ】Emme あわせ 2015/11/19 L)PEA+IOL 施行 2015/11/24 R)PEA+IOL 施行 術後経過良好のため 11/25 退院となった。	両眼	両眼	○
6	2015/11/30：△△眼科より紹介。白内障、緑内障。視力右 0.04(0.3)左 0.3(0.4)。カルテオロール 1 点中。 【予定】2/17 入院、2/18 右白内障、2/23 左白内障、2/24 退院 2016/02/18 R)PEA+IOL 施行（●●・■■Dr) no trouble, in the bag 術後視力 0.5（若い頃からこれくらいであった、と） 2016/02/23 L)PEA+IOL 施行（●●・■■Dr) no trouble, in the bag 術後経過良好で 2/24 退院となった 術後視力右(0.5)左(0.4) 次回、HFA。	両眼	片眼	×

誤判別表の中で誤判別の頻度が高いケース(頻度 3 以上)をピックアップし、表 6 に示した。「白内障」や「肺の悪性腫瘍」など同じ病状内で記載内容が似通ってしまっており、誤判別となっているが，詳細については，さらなる検証が必要となる。このあたりの原因が解明され，対策を行うことができれば，さらなる精度の向上が期待できる。

100 回試行の実行速度の比較を図 5 に示す。

ピークであったキーワード 250 に対して，決定木が 13 分，SVM が 2 時間 36 分，Random Forest が 3 時間 3 分であったのに対し，深層学習では 7 時間 24 分であった。今後，特に実行時間の長い深層学習は中間層を含めたパラメータの設定等で高速化を図れるかどうか検討する必要がある。

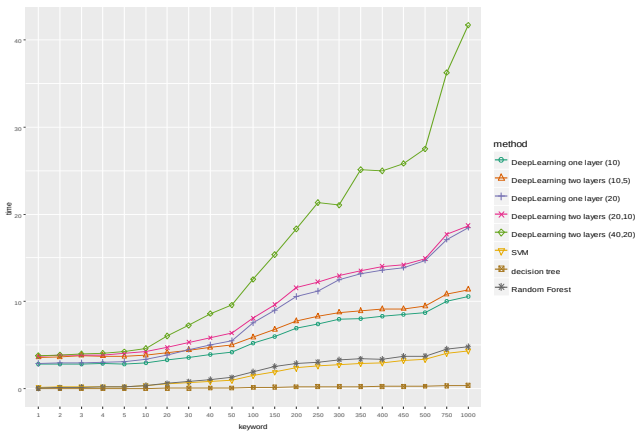

図 5　各手法のキーワード数別による実行速度

6. おわりに

本研究では，テキストに形態素解析を適用した後，対応分析による DPC 毎のキーワードの選定を行い，選ばれたキーワードに深層学習を適用し，キーワードによる DPC コーディングを行う分類器を構築した。

退院時サマリを用いた検証結果から，SVM，決定木が最大 78%程度の正答率であったのに対し，Random Forest が 90%以上の正答率を上げ，深層学習では 86%以上の精度を上げることができた。

本データでは実行速度，性能の点を考えれば，Random Forest を分類器構築方法として選択するのが最適であると考えられたが，深層学習についてはまだまだパラメータ調整の余地がある。

テキストマイニングにより診療文書内容から疾患の特定・分類などの可能性が示唆された。この結果を踏まえて，退院時サマリの記載内容から DPC コードが推定できるとすれば，退院患者に付与された DPC コードが正しいかどうかチェックを行うなど，実際に診療支援に役立てることが可能となると考える。

今後，誤判別を引き起こす抽出データの体系・構造を考慮に入れながら、対象症例を拡大させるなど，さらなる検証を行いたい。

参考文献

[1] Eric J Topol. High-performance medicine: the convergence of human and artificial intelligence. : Nature Medicine, 2019. : 44-56. vol. 25.

[2] 津本周作，矢入郁子，古崎晃司．医用人工知能研究会，人工知能 (特集: 研究会紹介) 34(5): 644-647, 2019.

[3] 三浦康秀，荒牧英治，大熊智子，外池昌嗣,杉原大悟，増市博，大江和彦: 電子カルテからの副作用関係の自動抽出．言語処理学会第 16 回年次大会，pp78-81，2010.

[4] 荒牧英治，三浦康秀，外池昌嗣，大熊智子増市博，大江和彦 ：退院サマリ文章可視化システムの構築．言語処理学会第 15 回年次大会，pp.348-351，2009.

[5] 小野大樹，高林克日己，鈴木隆弘，横井英人，井宮 淳，里村 洋一 ：テキストマイニングによる退院時要約自動分類の試み．医療情報学 24(1), p.35-44, 2004.

[6] F. Sebastiani.: Classification of text, automatic. In Keith Brown (ed.), The Encyclopedia of Language and Linguistics, Volume 14, 2nd Edition, Elsevier Science Publishers, Amsterdam, NL, 2006, pp. 457-462.

[7] Ji-Hyun Kim, Estimating classification error rate: Repeated cross-validation, repeated hold-out and bootstrap, Computational Statistics & Data Analysis, 53(11), p.3735-3745, 2009.

[8] InterSystems : InterSystems Caché, https://www.intersystems.com/jp/products/cache/ (2019)

[9] 石田基広 ：RMeCab, http://rmecab.jp/wiki/index.php?RMeCabFunctions (2016)

[10] Therneau, T.M. and Atkinson,E.J.: An Introduction to Recursive Partitioning Using the RPART Routines (2015)

[11] Karatzoglou, A., Smola, A., Hornik, K., and Zeileis, A.: kernlab -- An {S4} Package for Kernel Methods in {R}, Journal of Statistical Software,Vol.~11, No.~9, pp. 1-20 (2004)

[12] Liaw,~A. and Wiener,~M.: Classification and Regression by randomForest, R News, Vol.2, No.3, pp. 18-22 (2002)

[13] Drees,M.: Implementierung und Analyse von tiefen Architekturen in R, Master's thesis, Fachhochschule Dortmund (2013)

表 4　DeepLearning による分類正誤表

DeepLearning予測のDPCコード

正解のDPCコード

	010060x099030x	020110xx97xxx0	020110xx97xxx1	020160xx97xxx0	040040xx97x0xx	040040xx99 00xx	040040xx99 04xx	040040xx99 100x	040080x099 x0xx	060050xx97x0xx	060340xx03x00x	070560xx99x0xx	100070xxxxxxxx	110280xx99000x	12002xxx99x40x	120060xx02xxxx	120180xx01xxxx	130030xx99x40x	140010x199x00x	160620xx01xxxx
010060x099030x	13											1		3						
020110xx97xxx0		7	17								1									
020110xx97xxx1		25	41									1								
020160xx97xxx0		2		27																
040040xx97x0xx					5		1	1		1										
040040xx99 00xx						24	5		2					1						1
040040xx99 04xx						3	10	1							2			1		
040040xx99 100x						3	2	26										1		
040080x099 x0xx	2					2			11	2	1	7		3						
060050xx97x0xx			1			3				4	8		1					1		
060340xx03x00x	2	2	1	1					2	2	5									
070560xx99x0xx	1							2				6		8						
100070xxxxxxxx											1		13	2						
110280xx99000x	1	1				2			1	1			5	12			1			
12002xxx99x40x			1			1					1				3					
120060xx02xxxx																5				
120180xx01xxxx									2							2	3			
130030xx99x40x					1					1								20		
140010x199x00x																			3	
160620xx01xxxx																				1

表 6　DeepLearning の誤判別の頻度が高いケース

正解	予測	誤判別頻度
白内障、水晶体の疾患　手術手術あり　重症度等片眼 (020110xx97xxx0)	白内障、水晶体の疾患　手術手術あり　重症度等両眼 (020110xx97xxx1)	17
白内障、水晶体の疾患　手術手術あり　重症度等両眼 (020110xx97xxx1)	白内障、水晶体の疾患　手術手術あり　重症度等片眼 (020110xx97xxx0)	25
肺の悪性腫瘍　手術手術なし　処置等1あり　処置等2なし　副傷病名なし　(040040xx99100x)	肺の悪性腫瘍　手術手術なし　処置等1なし　処置等2なし (040040xx9900xx)	3
肺の悪性腫瘍　手術手術なし　処置等1なし　処置等24あり (040040xx9904xx)	肺の悪性腫瘍　手術手術なし　処置等1なし　処置等2なし (040040xx9900xx)	3
全身性臓器障害を伴う自己免疫性疾患　手術手術なし　処置等2なし　(070560xx99x0xx)	慢性腎炎症候群・慢性間質性腎炎・慢性腎不全　手術手術なし　処置等1なし　処置等2なし　副傷病名なし (110280xx99000x)	8
肺炎、急性気管支炎、急性細気管支炎（15 歳以上）手術手術なし　処置等2なし　(040080x099x0xx)	全身性臓器障害を伴う自己免疫性疾患　手術手術なし　処置等2なし　(070560xx99x0xx)	7
肺炎、急性気管支炎、急性細気管支炎（15 歳以上）手術手術なし　処置等2なし　(040080x099x0xx)	慢性腎炎症候群・慢性間質性腎炎・慢性腎不全　手術手術なし　処置等1なし　処置等2なし　副傷病名なし (110280xx99000x)	3
肺の悪性腫瘍　手術手術なし　処置等1なし　処置等2なし (040040xx9900xx)	肺の悪性腫瘍　手術手術なし　処置等1なし　処置等24あり (040040xx9904xx)	5
慢性腎炎症候群・慢性間質性腎炎・慢性腎不全　手術手術なし　処置等1なし　処置等2なし　副傷病名なし (110280xx99000x)	2型糖尿病（糖尿病性ケトアシドーシスを除く。） (100070xxxxxxxx)	5
肝・肝内胆管の悪性腫瘍（続発性を含む。）手術その他の手術あり　処置等2なし　(060050xx97x0xx)	肺の悪性腫瘍　手術手術なし　処置等1なし　処置等2なし (040040xx9900xx)	3
肝・肝内胆管の悪性腫瘍（続発性を含む。）手術その他の手術あり　処置等2なし　(060050xx97x0xx)	胆管（肝内外）結石、胆管炎　手術限局性腹腔膿瘍手術等　処置等2なし　副傷病名なし　(060340xx03x00x)	8
脳梗塞（JCS10 未満）手術手術なし　処置等1なし　処置等23あり　副傷病名なし　(010060x099030x)	慢性腎炎症候群・慢性間質性腎炎・慢性腎不全　手術手術なし　処置等1なし　処置等2なし　副傷病名なし (110280xx99000x)	3