The Review of Socionetwork Strategies (2021) 15:49–66
https://doi.org/10.1007/s12626-021-00076-7

ARTICLE

Determination of Disease from Discharge Summaries

A Text Mining Approach

Shusaku Tsumoto[1] · **Tomohirno Kimura**[2] · **Shoji Hirano**[1]

Received: 5 October 2020 / Accepted: 7 March 2021 / Published online: 5 May 2021

Abstract
Determining whether correct disease codes are included in discharge summaries is important for hospital management because submission of medical receipts with incorrect disease codes can result in loss of insurance reimbursement. Because medical information managers in large hospitals must evaluate more than 1000 summaries per month, an automated determination of discharge summaries will reduce their workload, allowing information managers to focus on complicated cases. This paper proposes a method of constructing classifiers of discharge summaries. In the first step, morphological analysis generated a term matrix from text data extracted from the hospital information system. Subsequently, important keywords were selected from an analysis of correspondence, training examples were generated, and machine learning methods were applied to the training examples. Several machine learning methods were compared using discharge summaries stored in the information system of Shimane University Hospital. A random forest method was found to be the best classifier when compared with deep learning, SVM and decision tree methods. Furthermore, the random forest method had a classification accuracy greater than 90%.

Keywords Discharge summary · Hospital information system · Classification learning · Random forest

This research was supported by a Grant-in-Aid for Scientific Research (B) 18H03289 from the Japan Society for the Promotion of Science (JSPS). On behalf of all authors, the corresponding author states that there are no conflicts of interest.

✉ Shusaku Tsumoto
tsumoto@med.shimane-u.ac.jp

Extended author information available on the last page of the article

1 Introduction

Computerization of patient records enables the storage of "big unstructured text data" in hospital information systems (HIS). For example, Shimane University Hospital treats about 1000 patients in its outpatient clinics and about 600 patients in inpatient wards. The HIS of this hospital stores about 200 GB of text data per year, including patient records, discharge summaries and radiology and pathology reports. Text mining of these resources can enable decisions about clinical actions, research and hospital management.

This paper proposes a five-step method of constructing classifiers for discharge summaries. In the first step, discharge summaries are extracted from the HIS. In the second step, morphological analysis is applied to a set of summaries and a term matrix is generated. In the third step, correspondence analysis is applied to the term matrix with class labels, and two-dimensional coordinates are assigned to each keyword; measurements of distances between categories and assigned points can generate a ranking of keywords for each category. In the fourth step, keywords are selected as attributes according to their rank, and training examples for classifiers are generated. Finally, learning methods are applied to the training examples. Experimental validation was performed using four methods: random forest, deep learning (multi-layer perceptron), SVM and decision tree induction. The random forest achieved the best performance, followed by the deep learning method.

The paper is organized as follows. Section 2 explains our motivation. Section 3 describes a proposed mining process. Section 4 shows the experimental results. Section 5 discusses these results. Finally, Sect. 6 provides the conclusions of this study.

2 Motivation

The principal purpose of applying AI to hospital data is to enhance the efficiency of the medical staff in a clinical environment. One of the more laborious tasks for doctors and nurses is documentation, including detailed descriptions of patient records. Careful documentation is needed for several purposes, including submission to insurance companies for reimbursement and exchange of information among hospitals and clinics. The accuracy of medical documents should be evaluated, mainly because most medical payments are based on the submission of medical fee statements, with the information from these statements obtained from medical documents. Large-scale hospitals in Japan must submit statements according to the Diagnosis Procedure Combination (DPC system) [4]. A DPC

code is assigned to the condition to which the majority of medical resources were devoted during the hospitalization of a patient. For each day of hospital stay, a payment point is assigned for each DPC code. Thus, medical payments by DPC code depend on the length of hospital stay, diagnosis and medical procedures, and differ from the traditional medical payment system, which depends on a set repayment for each medical procedure.[1]

Because DPC codes in the HIS are used to classify each medical payment during hospitalization, the assigned code may differ from medically classified diseases, making it difficult for doctors to assign DPC codes.

Thus, before submitting requests for payment, medical information managers must review clinical records to determine whether DPC codes are or are not correct. Mainly, managers check the validity of assigned DPC codes by reviewing discharge summaries and patient records. For example, at Shimane University Hospital, an average of 40 patients are discharged per day. In one month, about 1200 patients are discharged, which means that medical information managers must check 1200 discharge summaries and patient records per month, thus making efficient checking very important. At Shimane University Hospital, six managers check patient records and DPC codes.

An automated document classification system with correct DPC codes will help medical information managers at large hospitals submit accurate fee statements, enabling them to focus on complicated cases.

3 Methods

3.1 Discharge Summary

A discharge summary has been defined as a document that outlines the details of the hospitalization and care of a patient [1]. This summary is prepared when a patient is released from a health care facility and is incorporated into the permanent medical records of that patient. Ideally, a discharge summary should include an explanation for the patient's admission; records of patient complaints, physical findings, laboratory results and radiographic studies while hospitalized; a list of changes in medications at discharge; and recommendations for follow-up care. For optimal patient care, the discharge summary should be transmitted to or reviewed with the patient's primary care provider.

A discharge summary includes all clinical processes during patient hospitalization. It is written in more formal style than regular patient records. Thus, a conventional text mining approach can be used to extract enough keywords from the text of each discharge summary. Figure 1 shows an example of a discharge summary.

[1] Outpatient clinics utilize action-based payment systems, even in large hospitals.

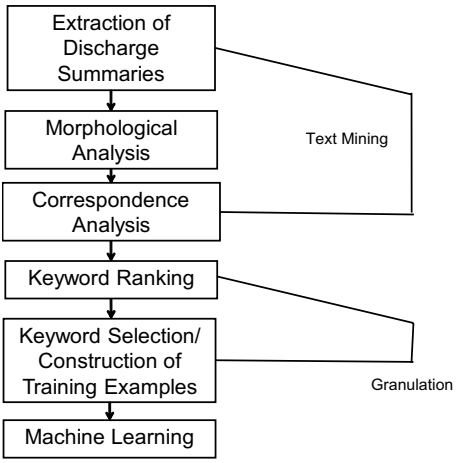
```
XXXX XXXX: 65 years old, Male
Chief complaints: nothing special

Present History:
Although high blood sugar (BS) was detected during a medical check up at his company,
the patient has not visited the clinic for 10 years. At age 61 years, his blood sugar
concentration had increased. He visited hospital A,
where he was diagnosed with diabetes mellitus, and prescribed MG500 mg.
Last August, he stopped his medication.

On December 15, 2015, he visited hospital B. His blood sugar was
168 mg/dl, and his HbA1c was 6.8\%.
 He was started on Grativ 50 mg/day. However, his HbA1c had not been lower than 7.9\%,
and he was scheduled for diabetes education.
He was admitted to  the University Hospital on February 22, 2016, for this purpose.
```

Fig. 1 An example of discharge summary

Fig. 2 Data mining process



3.2 Motivation for Feature Selection

Feature selection is important even for deep learners [11, 12]. Although deep learners show better performance in image analysis, differences between deep learners and other classification methods are generally very small. This may be due to the lack of a suitable network structure and the absence of suitable features for classification. Empirical results showing that deep learners are good at recognition of images suggest that some type of topological relationship should be explicitly embedded into training data. This study proposes a new feature selection method based on correspondence analysis, which calculates mapping attributes to points of multi-dimensional coordinates. The method can extract the topological relationships between keywords and data concepts.

3.3 Mining Process

Figure 2 shows the proposed total mining process, whose workflow consists of five steps.

3.3.1 Morphological Analysis

Target discharge summaries are extracted from the HIS, followed by morphological analysis [5], which outputs a term matrix, consisting of a contingency table for keywords and concepts.

3.3.2 Correspondence Analysis

The term matrix is subjected to correspondence analysis. Although high dimensional coordinates can be selected, a very large table is obtained. This study, therefore, focused on two-dimensional analysis, which can be easily used for visualization. Two dimensional coordinates are therefore assigned to each keyword and concept.

3.3.3 Ranking

The coordinates of each concept and keyword are used to calculate the euclidean distance between them. Distances are used to rank keywords to each concept, with smaller distances indicating a higher ranking.

3.3.4 Keyword Selection

Prior to analysis, the number of keywords is determined; e.g., 100. All keywords with rankings up that determination are selected for classification. Because some keywords may overlap, any overlapping keywords are deleted. Training examples with a classification label and the value of selected keywords (binary attributes) are subsequently constructed.

3.3.5 Classification

Finally, classification learning methods are applied. This study compared four classification methods: random forest [9], deep learning (multi-layer perceptron) (darch), Support Vector Machine (SVM) [7], Backpropagation Neural Network (BNN) [15] and decision tree (rpart) [14].

4 Experimental Evaluation

The 20 most frequent DPC codes in the fiscal year 2015 were selected, and discharge summaries that included these codes were extracted from the HIS of Shimane University Hospital. Table 1 shows the statistics of these 20 DPC codes, as well as the average number of characters used in the summaries.

Table 1 DPC codes of the top 20 diseases treated during fiscal year 2015

No	DPC	Cases
1	Cataracts (lateral)	445
2	Cataracts (bilateral)	152
3	Type II diabetes mellitus (except for keto-acidosis)	145
4	Lung cancer (with surgical operation)	131
5	Uterine cancer (without surgical operation)	121
6	Lung cancer (without surgical operation, chemotherapy)	120
7	Uterine benign tumor	111
8	Lung cancer (without surgical operation, with chemotherapy)	110
9	Miscarriage of pregnancy	110
10	Injury to the elbow and knee	99
11	Autoimmune disease	96
12	Non-Hodgkin disease	94
13	Pneumonia	86
14	Lung tumor (without surgical operation nor chemotherapy)	85
15	Chronic nephritis	83
16	Liver cancer	82
17	Gallbladder stone	82
18	Cerebral infarction	80
19	Retinal detachment	75
20	Fetal abnormalities	75

Except for extraction from data from the HIS, all processes were performed using R 3.5.0 software, with analysis and evaluation on two units HP Proliant ML110 Gen9 (Xeon E5-2640 v3.2 2.6GHz 8Core, 64GBDRAM) computers.

4.1 Mining Process

4.1.1 Correspondence Analysis

Morphological analysis was performed using RMeCab [5]. A bag of keywords was generated and used to construct a contingency table for these summaries. Correspondence analysis was applied to the table using the MASS package on R3.5.0. Two-dimensional coordinates were assigned to each keyword and each class.[2]

Figure 3 shows the two-dimensional plot of correspondence analysis. Because discharge summaries are written in English, all the keywords in the figure are shown in Japanese. English tranlations of some important keywords of frequent diseases

[2] The method can also generate $p(p \geq 3)$-dimensional coordinates. However, higher dimensional coordinates did not provide better performance than the experiments shown below.

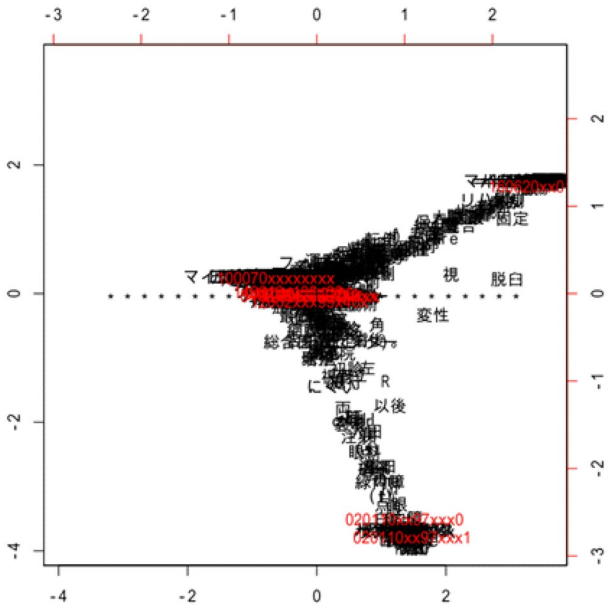


Fig. 3 Results of correspondence analysis

are shown in Table 3. All the keywords in Fig. 3 are arranged along a horseshoe curve, a specific feature of both correspondence analysis and principal component analysis [2, 3, 13]. These findings indicate that the correspondence analysis fit the correspondence between keywords and DPC codes.

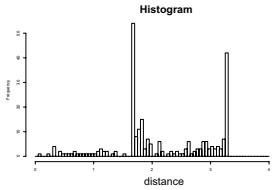
The information important for classification in Fig. 3 is plotted near the target classes, with the target class values (DPC codes) plotted as numerical codes.

For example, the two right bottom numbers denote a cataract, with the keywords for eye symptoms and surgical operations plotted near these classes. In contrast, the right upper class is "Injury to the Elbow and Knee", with the keywords for rehabilitation and fixation of joints located nearby.

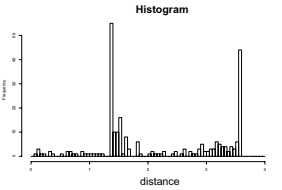
4.1.2 Ranking

Next, the distances between the coordinates of a keyword and those of a class are calculated, and the keywords ranked for each class. Because target classes and keywords are assigned to two dimensional planes, the distances between classes and keywords can be calculated from the assigned coordinates.

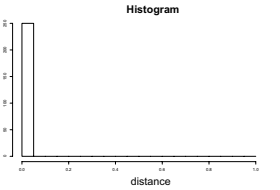
Figure 4 shows the distribution of the distances. The horizontal axis denotes the distance between keywords and classes and the vertical axis denotes the number of attributes of the given distance. Because distances close to 0 indicate that keywords and classes are very close, the figure shows that, except for cataracts and injury to the elbow and knee, the keywords are very close to the coordinates of each class. Thus, selection of keywords may be a little subtle and surrogate split may be useful for the decision tree induction and random forest methods.



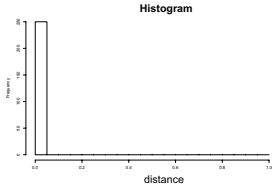
(a) Cataracts (Unilateral)



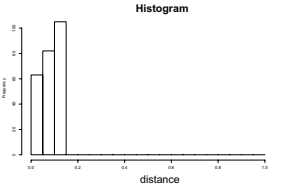
(b) Cataracts (Bilateral)



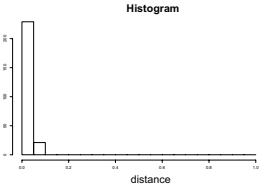
(c) Type II DM



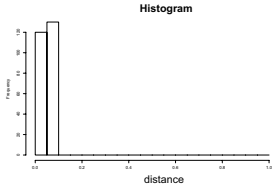
(d) Lung Cancer (with Surgical Operation)



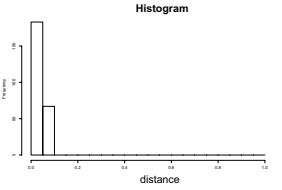
(e) Uterin Cancer



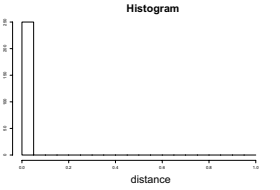
(f) Lung Cancer (without Surgical Operation, with Bronchoscopy)



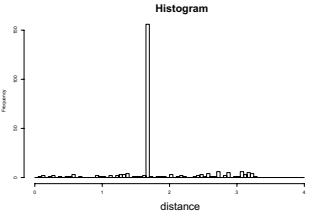
(g) Uterin Benign Tumor



(h) Lung Cancer (without Surgical Operation, with Chemotherapy)



(i) Miscarriage of Pregnancy



(j) Injury to the Elbow and Knee

Fig. 4 Distribution of distances between keywords and target classes for the top ten diseases and top 250 keywords. The horizontal axis denotes the distances between keywords and classes and the vertical axis denotes the number of attributes of the given distance. Because the keywords and classes are very close when the distances were nearly equal to 0, the figures shows that, except for cataracts and injury to the elbow and knee, the keywords were very close to the coordinates of each class

Table 2 Numbers of selected keywords and numbers of actually used keywords

#Keyword	Selected keywords	
	DPC Top 10	DPC Top 20
1	10	19
2	19	37
3	27	54
4	36	71
5	44	88
10	88	167
20	115	309
30	247	449
40	334	597
50	406	718
100	724	1125
150	1000	1472
200	1192	1782
250	1382	1932
300	1547	2031
350	1676	2113
400	1797	2192
450	1929	2273
500	2028	2364
750	2304	2808
1000	2545	3000
ALL	13944	20417

Using their rankings, a preset number of keywords was selected to generate a table for learning classification.

4.1.3 Keyword Selection

Table 2 shows the total numbers of keywords selected for the top 10 and 20 DPC codes. The selection of 250 keywords for each DPC code would result in a total of 5000 keywords. After the removal of overlapping keywords, only 1932 keywords were used for classification. Some important keywords may be deleted due to overlap if these keywords are frequently used in at least two diseases.

Table 3 shows the top 10 keywords for the three top DPC codes. For comparison, the results obtained by tf-idf are also shown. Interestingly, keywords selected by correspondence analysis differed from those selected by tf-idf, suggesting that frequency based information may not play important roles in the classification of discharge summaries.

Table 3 Top 10 keywords selected for each of the top three diseases (English translation)

Order	Cataracts(Uni)		Cataracts(Bil)		Type II-DM	
	Corres	TF-IDF	Corres	TF-IDF	Corres	TF-IDF
1	in	IOL	eye	IOL	keton	mg
2	VA	PEA	(+),	PEA	CTR	ml
3	lower	sunbetazone	mature	(Dr)	hr	hour
4	DC	left	abastine	+	Enhance	g
5	Eye_drops	+	Vitreous	Cataract	FDP	blood_sugar
6	Vitreous	right	Allergy	execute	noise	
7	mature	disclose	lower	trouble	jumping_pain	
8	abastine	vegamox	In	visual Eyesight	hypotensive	
growth	Lung					
9	(+),	Cataract	VA	discharge	pancreatitis	admission
10	#.	month	The	left	menopause	

Top 10 keywords are shown for three top diseases.

Corres: keywords selected by ranking of Correspondence Analysis.

TF-IDF: keywords selected by ranking of Tf-idf

4.1.4 Classification

Finally, decision tree (package: rpart [14]), random forest (package: randomForest [9], SVM (kernlab [7]), BNN(package: nnet [15]) and Deep Learner (multi-layer perceptron) (darch[3]) were applied to the generated training examples. To determine the parameters of Darch, the numbers of intermediate neurons were set at 20, (40,20) and (80,20), with an epoch of 100. For all other packages, the default settings of parameters were used.

4.1.5 Evaluation Process

The evaluation process was based on repeated two-fold cross validation [8].[4] First, the dataset was randomly split 1:1 into training samples and test samples. The training samples were used to construct a classifier, and the derived classifiers were evaluated with the remaining test samples. These procedures were repeated 100 times, and the averaged accuracy was calculated.

The number of keywords varied from 1 to 1000, selected according to their ranking by correspondence analysis.

[3] Darch was removed from R package. Please check the github: https://github.com/maddin79/darch.

[4] Two-fold cross-validation was selected because its estimator resulted in the lowest estimate of parameters, such as accuracy, as well as minimizing estimates of bias.

Table 4 Experimental results (averaged accuracies)

#keywords	Darch one layer (20)	Darch two layers (40,20)	Darch two layers (80,20)	SVM	Rpart	Random Forest	BNN
1	**0.247**	0.236	0.237	0.233	0.202	0.239	0.264
2	**0.442**	0.407	0.43	0.24	0.218	0.288	0.429
3	0.569	0.581	**0.584**	0.324	0.145	0.295	0.541
4	0.632	0.628	0.633	0.424	0.254	**0.676**	0.582
5	0.662	0.655	0.657	0.295	0.315	**0.714**	0.597
10	0.716	0.704	0.71	0.323	0.315	**0.767**	0.633
20	0.786	0.772	0.778	0.664	0.598	**0.826**	0.698
30	0.804	0.792	0.796	0.694	0.652	**0.841**	0.718
40	0.821	0.809	0.814	0.739	0.656	**0.855**	0.74
50	0.823	0.813	0.818	0.742	0.673	**0.855**	0.748
100	0.849	0.841	0.851	0.749	0.577	**0.875**	0.785
150	0.864	0.857	0.867	0.778	0.747	**0.896**	0.806
200	0.865	0.855	0.864	0.784	0.741	**0.907**	0.805
250	0.868	0.862	0.867	0.783	0.744	**0.906**	0.807
300	0.82	0.814	0.821	0.77	0.768	**0.907**	0.798
350	0.826	0.815	0.824	0.767	0.761	**0.907**	0.799
400	0.825	0.818	0.826	0.771	0.764	**0.908**	0.808
450	0.825	0.819	0.83	0.77	0.767	**0.908**	0.802
500	0.832	0.821	0.831	0.77	0.768	**0.908**	0.804
750	0.836	0.831	0.841	0.757	0.782	**0.907**	0.81
1000	0.846	0.836	0.845	0.753	0.79	**0.909**	0.82

All results are obtained by repeated two-fold cross validation (100 repetitions).

Layer(s) denote the number of intermediate layers.

and (a,b) shows the numbers of neurons for intermediate layers

4.2 Classification Results

Table 4 shows the evaluation results of the top 20 diseases. For four or fewer keywords, all the classifiers showed an accuracy of about 70%. At five or more keywords, however, SVM showed a decrease in accuracy, whereas the other methods showed monotonic increases in accuracy, with the latter plateauing at 200 keywords. The Random Forest method performed better than the other classifiers, followed by Darch deep learning. If more than 250 keywords were selected, the performance of Darch decreased, whereas the performances of random forest and decision trees increased monotonically. Although BNN showed poorer accuracy than Darch (default setting) with 5 to 100 selected keywords, the accuracy of BNN approached that of Darch classifiers with a larger number of keywords. Interestingly, the accuracy of the decision tree method increased monotonically, becoming maximal when all the keywords were used for analysis.

True \ Prediction	Brain infarction	Cataract (Uni)	Cataract (Bil)	Retina Detach-ment (Uni)	Lung Cancer (with Surgical Operation)	Lung Cancer (without Surgery)	Lung Cancer (with Broncho-scopy)	Lung Cancer (with Chemo-therapy)	Pneu-monia	Liver Cancer (with Surgical Operation)	Gold-bladder Stone & Infla-mmation	Severe Auto-immune Disease	Type II DB	Chronic Nephritis	Uterus Cancer	Uterus Benign Tumor	Fetal Ab-normali-ties	Non-hodgikin	Preg-nancy (early)	Injury of Knee or Elbow	Total
brain infarction	17																				17
Cataract (Uni)		20	5																		25
Cataract (Bil)		33	34																		67
Retina Detachment (Uni)		1		28																	29
Lung Cancer (with Surgical Operation)					6		1		1												8
Lung Cancer (without Surgery)						16	10	3	2					2							33
Lung Cancer (with Bronchoscopy)						1	16														17
Lung Cancer (with Chemotheory)							2	29	1												32
Pneumonia	1	3		1		1	1		17		2		1	1							28
Liver Cancer (with Surgical Operation)								1	1	11	4		1								18
Goldbladder Stone & Inflammation		5								1	9										15
Severe Autoimmune Disease								2	1			13		1							17
Type II DB										1			15								16
Chronic Nephritis	2	2							1			1	4	14							24
Uterus Cancer															6						6
Uterus Benign Tumor		1														4					5
Fetal Abnormalities																	7				7
Nonhodgikin								1										21			22
Pregnancy (early)																			3		3
Injury of Knee or Elbow														1							1
Total	20	65	39	29	6	18	30	36	24	13	15	14	21	19	6	4	7	21	3	0	390

Fig. 5 Confusion matrix obtained with the random forest method

True \ Prediction	Brain infarction	Cataract (Uni)	Cataract (Bil)	Retina Detach-ment (Uni)	Lung Cancer (with Surgical Operation)	Lung Cancer (without Surgery)	Lung Cancer (with Broncho-scopy)	Lung Cancer (with Chemo-therapy)	Pneu-monia	Liver Cancer (with Surgical Operation)	Gold-bladder Stone & Infla-mmation	Severe Auto-immune Disease	Type II DB	Chronic Nephritis	Uterus Cancer	Uterus Benign Tumor	Fetal Ab-normal-ities	Non-hodgikin	Preg-nancy (early)	Injury of Knee or Elbow	Total
Brain infarction	13											1		3							17
Cataract (Uni)		7	17								1										25
Cataract (Bil)		25	41									1									67
Retina Detachment (Uni)		2		27																	29
Lung Cancer (with Surgical Operation)					5		1	1		1											8
Lung Cancer (without Surgery)						24	5		2					1						1	33
Lung Cancer (with Bronchoscopy)						3	10	1							2			1			17
Lung Cancer (with Chemotheory)						3	2	26										1			32
Pneumonia	2					2			11	2	1	7		3							28
Liver Cancer (with Surgical Operation)			1			3				4	8		1					1			18
Goldbladder Stone & Inflammation	2	2	1	1					2	2	5										15
Severe Autoimmune Disease	1							2				6		8							17
Type II DB											1		13	2							16
Chronic Nephritis	1	1				2			1	1			5	12			1				24
Uterus Cancer			1			1					1				3						6
Uterus Benign Tumor																5					5
Fetal Abnormalities									2						2		3				7
Nonhodgikin					1					1								20			22
Pregnancy (early)																			3		3
Injury of Knee or Elbow																				1	1
Total	19	37	61	28	6	38	18	30	18	11	17	15	19	29	5	7	4	23	3	2	390

Fig. 6 Confusion matrix obtained with the deep learner method

5 Discussion

5.1 Misclassified Cases

Figures 5 and 6 show confusion matrices of random forest and darch (multi-layer perceptron), where DPC codes were set in order, indicating that similar codes were to similar diseases.[5] Shaded regions indicate misclassified patients. Although errors using the random forest method are located near the diagonal, errors using darch were more scattered. This finding suggests that the random forest method was almost correct in classifying a patient if similar DPC codes were grouped into one generalized class. In contrast, the Darch method had unexpected errors.

5.2 Classification Accuracy of Decision Trees

Two results were unexpected: (1) the accuracy of decision trees increased monotonically, and (2) the random forest method was more accurate than the other methods. Because the random forest method can be considered a refinement of decision tree induction, representation by decision trees may provide insight into hidden structures present in the discharge summaries.

Figure 7 shows the decision tree obtained with 1000 selected keywords extracted by morphological analysis, with 23 attributes used for description. Because the shape of the tree cannot be determined by linear combination, SVM, or linear combination of keywords, it may not show classification accuracy. Second, the selection process based on correspondence analysis may not be appropriate in selecting keywords for SVM.

Figure 8 shows the location of each keyword used in the decision tree based on its ranking in each classification class. All of the keywords were not selected by ranking, perhaps because the differences in distances among the attributes were very small. Future studies are needed to assess the nature of ranking.

A review of the decision tree by medical experts found that the tree was very compact but reasonable and that the selection of keywords was very interesting and explainable. This selection may reflect the differences in the description of disease summaries among the target diseases. Further evaluation should include a detailed examination of discharge summaries.

5.3 Execution Time

Two units of HP Proliant ML110 Gen9 (Xeon E5-2640 v3.2 2.6GHz 8Core, 64GBDRAM) workstation were used.

[5] DPC codes are a three-level hierarchical system, with each DPC code defined as a tree. The first level denotes the type of disease, the second level denotes the primary treatment selected for that patient, and the third-level shows any additional therapy. Thus, in the tables, characteristics of codes were representative of similarities.

```
[1] 0.8279387 ,n= 1210
node), split, n, loss, yval, (yprob)  * denotes terminal node
      1) root 1210 1005 Cataract (Uni) (0.072 0.032 0.046 0.056 0.039 0.027 0.027 0.044
                                        0.036 0.06 0.048 0.049 0.036 0.036 0.17 0.072 0.039 0.046 0.033 0.031)
        2) IOL< 0.5 893  807 Type2 DM (0.096 0.044 0.063 0.076 0.053 0.037 0.037 0.059 0.049
                                       0.082 0.065 0.066 0.048 0.049 0.0011 0 0.053 0.063 0.045 0.015)
          4) Fracture >=2.5 87    8  Type2 DM(0.91 0 0 0 0 0 0 0 0.011
                                     0.011 0.011 0 0.011 0 0 0 0.011 0.023 0.011 0) *
          5) Fracture < 2.5 806  734 Lung Cancer (with Surgical Operations)(0.0087 0.048 0.069
        0.084 0.058 0.041 0.041 0.066 0.053 0.089 0.071 0.073 0.052 0.055 0.0012 0 0.057 0.067 0.048 0.016)
            10) ics< 0.5 739  671 Uterus Cancer (with Chemotherapy) (0.0095 0.053 0.076 0.092
        0.064 0.045 0.045 0.072 0.058 0.0068 0.077 0.08 0.057 0.06 0.0014 0 0.062 0.073 0.053 0.018)
              20) TC>=0.5 66    3  Uterus Cancer (with Cehmotheray) (0 0 0 0.95 0 0 0 0 0.015 0 0 0.03
        0 0 0 0 0 0 0 0) *
              21) TC< 0.5 673  616  Lung Cancer (with Surgical Operations) (0.01 0.058 0.083 0.0074
        0.07 0.049 0.049 0.079 0.062 0.0074 0.085 0.085 0.062 0.065 0.0015 0 0.068 0.08 0.058 0.019)
                42) ROM< 0.5 620  563 Lung Cancer (with Surgical Operations)  (0.011 0.063 0.09 0.0081
        0.076 0.053 0.053 0.085 0.068 0.0081 0.092 0.092 0.068 0.071 0.0016 0 0.074 0.0016 0.063 0.021)
                  84) myoma >=1.5 57    4  Uterus Benign Tumor  (0.018 0 0.93 0 0.018 0.018 0 0.018
        0 0 0 0 0 0 0 0 0 0 0 0) *
                  85) myoma < 1.5 563  506 Lung Cancer (with Surgical Operation)  (0.011 0.069 0.0053
         0.0089 0.082 0.057 0.059 0.092 0.075 0.0089 0.1 0.1 0.075 0.078 0.0018 0 0.082 0.0018 0.069 0.023)
                   170) Fetus  >=0.5 50    3 Pregnancy  (0 0 0 0 0 0.06 0 0.94 0 0 0 0 0 0 0 0 0 0 0 0) *
                   171) Fetus < 0.5 513  456  Lung Cancer (with Surgical Operation)  (0.012 0.076 0.0058 0.0097
        0.09 0.057 0.064 0.0097 0.082 0.0097 0.11 0.11 0.082 0.086 0.0019 0 0.09 0.0019 0.076 0.025)
                     342) Lymphoma < 0.5 461  404 Lung Cancer (Chemotherapy) (0.011 0.082 0.0065 0.011
        0.098 0.061 0.072 0.011 0.091 0.011 0.12 0.12 0.089 0.091 0.0022 0 0.0043 0.0022 0.085 0.028)
                       684) TACE>=0.5 33    0 Liver Cancer (0 1 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0) *
                       685) TACE< 0.5 428  371 Lung Cancer (Chemotherapy)  (0.012 0.012 0.007 0.012
        0.11 0.065 0.077 0.012 0.098 0.012 0.13 0.13 0.096 0.098 0.0023 0 0.0047 0.0023 0.091 0.03)
                         1370) Barre>=0.5 36    2 Brain infarction (0 0 0 0 0.028 0 0 0 0.94 0 0 0.028 0 0 0 0 0 0 0 0) *
                         1371) Barre< 0.5 392  336 Lung Cancer (with Surgical Operation) (0.013 0.013 0.0077
        0.013 0.11 0.071 0.084 0.013 0.02 0.013 0.14 0.14 0.1 0.11 0.0026 0 0.0051 0.0026 0.099 0.033)
                           2742) IVCY>=0.5 30    0 Autoimmune Disease (0 0 0 0 1 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0) *
                           2743) IVCY< 0.5 362  306  Lung Cancer (with Surgical Operation)  (0.014 0.014 0.0083
        0.014 0.039 0.077 0.091 0.014 0.022 0.014 0.15 0.15 0.11 0.12 0.0028 0 0.0055 0.0028 0.11 0.036)
                             5486) EBUS>=0.5 32    0 Lung Cancer (with Surgical Operation)
                                                  (0 0 0 0 0 0 0 0 0 0 1 0 0 0 0 0 0 0 0 0) *
                             5487) EBU S< 0.5 330  274 Lun Cancer (Chemotherapy)
                                                    (0.015 0.015 0.0091 0.015 0.042 0.085 0.1 0.015 0.024 0.015
                                                     0.073 0.17 0.12 0.13 0.003 0 0.0061 0.003 0.12 0.039)

                              10974) Caesarean>=0.5 31    3  Fetal Abnormalities  (0 0 0 0 0 0.9 0 0.097 0 0 0 0 0 0 0 0 0 0 0 0) *
                              10975) Caesarean < 0.5 299  243 Lung Cancer (Chemotherapy)  (0.017 0.017 0.01
         0.017 0.047 0 0.11 0.0067 0.027 0.017 0.08 0.19 0.14 0.14 0.0033 0 0.0067 0.0033 0.13 0.043)
                                21950) Galdbladder >=1.5 38    7  Galdbladder Stone  (0 0.053 0 0 0.053 0 0.82
         0 0 0 0 0 0.026 0.053 0 0 0 0 0 0) *
                                21951) Galdbladder < 1.5 261  205  Lung Cancer (Chemotherapy)  (0.019 0.011 0.011
        0.019 0.046 0 0.0077 0.0077 0.031 0.019 0.092 0.21 0.15 0.15 0.0038 0 0.0077 0.0038 0.15 0.05)
                                  43902) CDDP>=0.5 48    9 Lung Cancer (Chemotherapy)  (0 0 0 0 0 0 0 0 0 0.021
        0.021 0.81 0.12 0.021 0 0 0 0 0 0) *
                                  43903) CDDP< 0.5 213  174 Pneumonia  (0.023 0.014 0.014 0.023 0.056 0 0.0094
        0.0094 0.038 0.019 0.11 0.08 0.16 0.18 0.0047 0 0.0094 0.0047 0.18 0.061)
                                    87806) Pneumonia < 2.5 174  136 Chronic Nephritis  (0.029 0.017 0.017 0.029
        0.046 0 0.011 0.011 0.046 0.017 0.13 0.098 0.18 0.057 0.0057 0 0.011 0.0057 0.22 0.075)
                                     175612) Kidney < 1.5 125   97 Lung Cancer (0.016 0.016 0.024 0.008 0.032 0
         0.016 0.016 0.064 0.024 0.18 0.14 0.22 0.072 0.008 0 0.008 0.008 0.048 0.1)
                                       351224) Fissure < 0.5 111   83 Lung Cancer  (0.009 0.018 0.027 0.009 0.036 0 0.018
        0.018 0.072 0.027 0.2 0.15 0.25 0.081 0.009 0 0.009 0.009 0.054 0)
                                         702448) VP< 0.5 99   71  Lung Cancer   (0.01 0.02 0.03 0.01 0.04 0 0.02 0.02
        0.081 0.03 0.22 0.051 0.28 0.091 0.01 0 0.01 0.01 0.061 0)
                                           1404896) Broncho >=2.5 18    3  Lung Cancer  (0 0 0 0 0.056 0 0 0 0 0
         0.83 0 0 0.11 0 0 0 0 0 0) *
                                           1404897) Broncho < 2.5 81   53  Lung Cancer  (0.012 0.025 0.037 0.012
        0.037 0 0.025 0.025 0.099 0.037 0.086 0.062 0.35 0.086 0.012 0 0.012 0.012 0.074 0) *
                                         702449) VP>=0.5 12    0 Lung Cancer (Chemotherapy)  (0 0 0 0 0 0 0 0 0 0 0 1
         0 0 0 0 0 0 0 0) *
                                       351225) Fissure>=0.5 14    1 Retinal Detachment (Uni)  (0.071 0 0 0 0 0 0 0 0
         0 0 0 0 0 0 0 0 0 0 0 0 0.93) *
                                     175613) Kidney >=1.5 49   17 Chrnoic Nephritis  (0.061 0.02 0 0.082 0.082 0 0
         0 0 0 0 0 0.061 0.02 0 0 0.02 0 0.65 0) *
                                    87807) Pneumonia >=2.5 39   10  (0 0 0 0 0.1 0 0 0 0 0.026 0.026 0 0.077
       0.74 0 0 0 0 0.026 0) *
                     343) Lymphoma>=0.5 52    8 NonHodgkin Lymphoma (0.019 0.019 0 0 0.019 0.019
         0 0 0 0 0.019 0 0.019 0.038 0 0 0.85 0 0 0) *
                  43) ROM>=0.5 53    0 Injury of Knee or Elbow  (0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 1 0 0) *
              11) ics>=0.5 67    0 Lung Cancer (with Surgical Operation)  (0 0 0 0 0 0 0 0 0 1 0 0 0 0 0 0 0 0 0 0) *
            3) IOL>=0.5 317  113 Cataract (Uni) (0.0032 0 0 0 0 0 0 0 0 0 0 0 0 0 0.64 0.27 0 0 0 0.079)
              6) PEA>=1.5 114   44  Cataract (Bil)  (0.0088 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0.36 0.61 0 0 0 0.018)
                12) This_time>=0.5 27    5 Cataracta (Uni) (0.037 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0.81 0.15 0 0 0 0) *
                13) This_time < 0.5 87   21 Cataract (Bil) (0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0.22 0.76 0 0 0 0.023) *
              7) PEA< 1.5 203   40 Cataract(Uni) (0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0.8 0.084 0 0 0 0.11)
                14) SF< 0.5 181   19 Cataract (Uni) (0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0.9 0.094 0 0 0 0.011) *
                15) SF>=0.5 22    1 Retinal Detachment (Uni) (0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0.045 0 0 0 0 0.95) *
```

Fig. 7 Decision tree with 1000 keywords from the top 20 diseases

Word	Cataract (Uni)	Cataract (Bil)	Type II-DM	Lung Cancer (with Surgical Operation)	Uterus Cancer (Chemotherapy)	Lung Cancer(Chemotherapy)	Uterus Benign Tumor	Lung Cancer (Broncho)	Pregnancy	Injury of Knee, Elbow	Autoimmune Disease	Non Hodgkin Disease	Pneumonia	Lung Cancer (without Surgical Operation nor Chemotherapy)	Chronic Nephritis	Liver Cancer	Galdbladdar Stone , Inflammation	Brain Infarction	Retinal Detachment	Fetal Abnormalities
-																				
IOL	120	112																	116	
Fracture			837																	
ics	513	514		142	330		242			540									514	294
TC	371	365		609	27	968	605	789						743		490		886	373	614
TROM																				
myoma	419	420		329	759		4			445									420	95
Pregnancy						595		613	562		223	530	167	626	462	748	311	628		
Lymphoma						886		716	885		895	704	646	893	610		829			
TACE	926	925		762	516	131	794	26	132		687	56	490	59	153	117	467	186	929	796
Barre				954	806	295	970	314	292		127	301	251	316	336	384	73	341		978
IVCY			455																	
EBUS					930	463		443	422		269	270	42	476	237	603	218	484		
Caesear	464	465		252	621		123			480									463	169
Galdbladdar			719								822						947			
CDDP					899	757		280	771			476	683	611	322	800	861	834		
Pneumonia			684								947									
Kidney			34																	
Fissure	28	18																	23	
VP						841		536	846			619	721	805	560	877	917	966		
Broncho				698	662	56	728	215	32	990	341	201	530	176	511	229	147	87		737
Pneumonia			684								947									
PEA	124	116																	119	
This_time	191	191								977									191	
PEA	124	116																	119	
SF	110	102																	106	

Fig. 8 Ranking of keywords in the decision tree for the top 20 diseases

Table 5 Times required for construction of classifications for the top 20 diseases

#keyword	Darch one layer (20)	Darch two layers (40, 20)	Darch two layers (80, 20)	SVM	Rpart	Random forest	BNN
1	172	226	230	8	0	3	3
2	175	231	247	10	0	6	4
3	177	239	257	12	1	9	4
4	182	245	276	12	0	10	5
5	186	254	288	14	1	11	6
10	201	277	368	21	1	20	11
20	231	362	532	31	2	35	42
30	269	435	729	40	3	48	95
40	302	516	891	48	4	61	171
50	331	575	1063	57	4	76	239
100	453	752	1574	90	7	115	577
150	540	922	1948	114	9	152	943
200	635	1099	2431	142	12	172	1429
250	672	1281	2902	156	13	183	1679
300	751	1263	2701	164	13	198	1850
350	789	1507	3085	172	14	204	2010
400	817	1500	3114	179	15	202	2136
450	833	1550	3311	194	15	222	2296
500	883	1650	3381	201	17	223	2509
750	1027	2177	4802	244	19	271	3542
1000	1110	2504	5030	261	22	288	4062

All results are obtained by repeated two-fold cross validation (100 repetitions).

Layer(s) denote the number of intermediate layers and (a, b) shows the numbers of neurons for intermediate layers

Table 6 Comparison of accuracies (tf-idf)

#keywords	Darch one layer (20)	SVM	Rpart	Random forest
1	0.113	0.507	0.09	0.565
2	0.165	0.585	0.074	0.638
3	0.194	0.594	0.183	0.708
4	0.195	0.605	0.311	0.788
5	0.197	0.611	0.365	0.804
10	0.23	0.65	0.56	0.839
20	0.199	0.67	0.769	0.882
30	0.195	0.678	0.778	0.895
40	0.193	0.685	0.802	0.903
50	0.18	0.681	0.802	0.907
100	0.194	0.686	0.8	0.915
150	0.206	0.685	0.797	0.916
200	0.203	0.683	0.796	0.914
250	0.212	0.682	0.795	0.913
300	0.19	0.683	0.802	0.912
350	0.217	0.683	0.802	0.911
400	0.202	0.679	0.801	0.91
450	0.168	0.681	0.803	0.909
500	0.187	0.68	0.802	0.908
750	0.191	0.678	0.802	0.904
1000	0.203	0.677	0.802	0.902

All results are obtained by repeated twofold cross validation (100 repetitions).

Layer(s) denote the number of intermediate layers (a, b) shows the numbers of neurons for intermediate layers

Table 5 shows an empirical comparison of repeated twofold cross validations (100 trials). The times need for Random Forest and SVM were 183 and 156 mins for 250 keywords, whereas Darch (20) required 672 mins. For 1000 keywords, the times needed for Random Forest, SVM and Darch (20) were 261, 288, and 1101 mins, respectively. The times required by random forest and BNN methods were close to those of Deep Learners. In the case of Darch, the number of intermediate layers resulted in greater computation times, although the growth rate was smaller than that of BNN.

5.4 Comparison with tf-idf

A major approach in text classification is ranking with tf-idf [6, 10]. Thus, tf-idf ranking was compared with the above approach using the same scheme as in Sect. 4. Interestingly, deep learning with tf-idf ranking showed much poorer performance than ranking by correspondence analysis, whereas random forest with tf-idf ranking was slightly better than that with ranking by correspondence

analysis for <200 keywords (Table 6). Because the average accuracy of deep learning different by only a few percent from that of the random forest method, ranking by correspondence analysis is a better approach for text classification by deep learning, at least in this applied domain. Ranking by correspondence analysis includes geometric information about keywords and concepts, with embedding of geometric knowledge being important for deep learning.

A major approach in text classification is ranking with tf-ifd, which were introduced by Luhn [10] and Sparck Jones [6]. Thus, here, we compared tf-idf ranking with the above approach by using the same scheme as in Sect. 4.

Interestingly, deep learning with tf-idf ranking performs much worse than that with ranking by correspondence analysis, whereas random forest with tf-idf ranking is a little better than that with ranking by correspondence analysis when the number of selected keywords is smaller than 200. These results are clearly shown in Table 6.

6 Conclusion

This study proposes a five-step method for constructing classifiers for discharge summaries. In the first step, discharge summaries are obtained from the HIS. In the second step, morphological analysis is applied to a set of summaries to generate a term matrix. In the third step, correspondence analysis is applied to the classification labels and term matrix, generating two-dimensional coordinates. Measurements of the distances between categories and assigned points enables ranking of keywords. In the fourth step, keywords are selected as attributes according to rank, and training examples for classifiers are generated. Finally, learning methods are applied to the training examples. This method was experimentally validated using discharge summaries from Shimane University Hospital during the 2015 fiscal year. Optimal performance was provided by the random forest method, with a classification accuracy of about 93%, followed by deep learning with a classification accuracy of about 91%. In contrast, decision tree methods with many keywords was slightly less accurate than neural networks and deep learning methods. The selected keywords and tree structure were deemed reasonable by domain experts, perhaps because the hidden structure of knowledge in a dataset may be close to the structure approximated by a set of trees and because deep learning may generate such structures in the networks. Our future work will attempt to validate this hypothesis.

Acknowledgements This research was supported by a Grant-in-Aid for Scientific Research (B) 18H03289 from the Japan Society for the Promotion of Science(JSPS).

Declarations

Conflict of interest On behalf of all authors, the corresponding author states that there are no conflicts of interest.

References

1. Discharge summary, http://medical-dictionary.thefreedictionary.com/discharge+summary. Accessed Feb 14, 2021
2. Deáth, G. (1999). Principal curves: A new technique for indirect and direct gradient analysis. *Ecology*, *80*(7), 2237–2253.
3. Hastie, T., & Stuetzle, W. (1989). Principal curves. *Journal of the American Statistical Association*, *84*(406), 502–516.
4. IgakuTsushinsha (ed.) (2020). Quick Reference of DPC points (in Japanese). IgakuTsushinsha, Tokyo
5. Ishida, M. (2016). Rmecab. http://rmecab.jp/wiki/index.php?RMeCabFunctions
6. JONES, K. S. (1972). A statistical interpretation of term specificity and its application in retrieval. *Journal of Documentation*, *28*(1), 11–21.
7. Karatzoglou, A., Smola, A., Hornik, K., & Zeileis, A. (2004). kernlab - an S4 package for kernel methods in R. *Journal of Statistical Software, 11*(9), 1–20. http://www.jstatsoft.org/v11/i09/
8. Kim, J. H. (2009). Estimating classification error rate: Repeated cross-validation, repeated hold-out and bootstrap. *Computational Statistics and Data Analysis*, *53*(11), 3735–3745. https://doi.org/10.1016/j.csda.2009.04.009.
9. Liaw, A., & Wiener, M. (2002). Classification and regression by randomforest. *R News, 2*(3), 18–22. http://CRAN.R-project.org/doc/Rnews/
10. Luhn, H. P. (1957). A statistical approach to mechanized encoding and searching of literary information. *IBM Journal of Research and Development*, *1*(4), 309–317.
11. Mares, M. A., Wang, S., & Guo, Y. (2016). Combining multiple feature selection methods and deep learning for high-dimensional data. *Transactions on Machine Learning and Data Mining*, *9*, 27–45.
12. Nezhad, M. Z., Zhu, D., Li, X., Yang, K., & Levy, P. (2017). SAFS: A deep feature selection approach for precision medicine. arXiv:1704.05960
13. Podani, J., & Miklós, I. (2002). Resemblance coefficients and the horseshoe effect in principal coordinates analysis. *Ecology*, *83*(12), 3331–3343.
14. Therneau, T. M., & Atkinson, E. J. (2015). An Introduction to Recursive Partitioning Using the RPART Routines. https://cran.r-project.org/web/packages/rpart/vignettes/longintro.pdf
15. Venables, W. N., & Ripley, B. D. (2002). Modern Applied Statistics with S. Springer, New York, 4th edn. http://www.stats.ox.ac.uk/pub/MASS4, iSBN 0-387-95457-0

Authors and Affiliations

Shusaku Tsumoto[1] · Tomohirno Kimura[2] · Shoji Hirano[1]

Tomohirno Kimura
t-kimura@med.shimane-u.ac.jp

Shoji Hirano
hirano@med.shimane-u.ac.jp

[1] Present Address: Department of Medical Informatics, Faculty of Medicine, Shimane University, 89-1 Enya-cho, Izumo 693-8501, Japan

[2] Medical Services Division, Faculty of Medicine, Shimane University, 89-1 Enya-cho, Izumo 693-8501, Japan