



HAL
open science

From Texts to Classification Knowledge

Shusaku Tsumoto, Tomohiro Kimura, Shoji Hirano

► **To cite this version:**

Shusaku Tsumoto, Tomohiro Kimura, Shoji Hirano. From Texts to Classification Knowledge. 4th International Conference on Intelligence Science (ICIS), Feb 2021, Durgapur, India. pp.173-184, 10.1007/978-3-030-74826-5_15 . hal-03741713

HAL Id: hal-03741713

<https://inria.hal.science/hal-03741713v1>

Submitted on 1 Aug 2022

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.



Distributed under a Creative Commons Attribution 4.0 International License

From Texts to Classification Knowledge ^{*}

Shusaku Tsumoto¹, Tomohiro Kimura², and Shoji Hirano¹

¹ Department of Medical Informatics, Faculty of Medicine, Shimane University
{tsumoto, hirano}@med.shimane-u.ac.jp

² Medical Services Division, Faculty of Medicine, Shimane University
t-kimura@med.shimane-u.ac.jp

Abstract. Hospital information system stores all clinical information, whose major part is electronic patient records written by doctors, nurses and other medical staff. Since records are described by medical experts, they are rich in knowledge about medical decision making. This paper proposes an approach to extract clinical knowledge from the texts of clinical records. The method consists of the following three steps. First, discharge summaries, which include all clinical processes during the hospitalization, are extracted from hospital information system. Second, morphological and correspondence analysis generates a term matrix from text data. Then, finally, machine learning methods are applied to a term matrix in order to acquire classification knowledge. We compared several machine learning methods by using discharge summaries stored in hospital information system. The experimental results show that random forest is the best classifier, compared with deep learning, SVM and decision tree. Furthermore, random forest gains more than 90% classification accuracy.

Keywords: Discharge Summary · Hospital Information System · Text Mining · Classification Learning · Rough Sets

1 Introduction

Computerization of patient records enables to store “big unstructured text data” in a Hospital Information System (HIS). For example, our system in Shimane university hospital, where about 1000 patients visit outpatient and about 600 patients stays inpatient, additionally stores about 200GB text data per year, including patient records, discharge summaries and reports of radiology and pathology. Application of text mining to these resources is very important to discover useful knowledge from text data, which can be viewed as a new type of support of clinical actions, researches and hospital management.

This paper is a first step to acquire clinical knowledge from texts: We focus on discharge summaries, which includes all clinical processes during the hospitalization, whose written style is formal, compared with regular records. Thus, conventional text mining approach can be used to extract enough keywords from each text. Then, we

^{*} This research is supported by Grant-in-Aid for Scientific Research (B) 18H03289 from Japan Society for the Promotion of Science(JSPS).

propose a method for construction of classifiers of diseases codes from discharge summaries, which consists of the following five steps. First, discharge summaries are extracted from hospital information system. Second, morphological analysis is applied to a set of summaries and a term matrix is generated. Second, correspond analysis is applied to the term matrix with the class labels, and two dimensional coordinates are assigned to each keyword. By measuring the distance between categories and the assigned points, ranking of key words for each category will be generated. Then, keywords are selected as attributes according to the rank, and training example for classifiers will be generated. Finally learning methods are applied to the training examples. We conduct experimental validation with four methods, random forest, deep learning (multi-layer perceptron), SVM and decision tree induction method. The results show that random forest achieved the best performance and the second best was the deep learner with a small difference, but decision tree methods with many keywords performed only a little worse than neural network or deep learning methods. The paper is organized as follows. Section 2 explains our motivation. Section 3 gives a proposed mining process. Section 4 shows the experimental results. Section 5 discusses the results obtained. Finally, Section 6 concludes this paper.

2 Motivation

Knowledge acquisition from medical experts is a very classical problem, but still a bottleneck for developing medical expert systems [10, 12].

The reason why knowledge acquisition is difficult is that medical experts had difficulties in reformatting their knowledge into if-then rules. Thus, although deeper reasoning for experts emerged in the research of artificial intelligence in medicine in order to solve such bottle neck problems [1, 11], it is still a difficult problem.

However, medical staff write down the process of clinical decision making as texts: it suggests that such knowledge about decision making can be represented as language texts, but not as rule-based. Thus, if we can analyze texts written by medical staff, we will be able to get knowledge about medical decision making.

3 Methods

3.1 Discharge Summary

In this paper, we focus on discharge summaries. A discharge summary includes all clinical processes during the hospitalization, whose written style is formal, compared with regular records. Thus, conventional text mining approach can be used to extract enough keywords from each text. We apply ordinary text mining process as preprocessing.

3.2 Motivation for Feature Selection

It is well known that feature selection is important even for deep learners [8, 9]. Although deep learners gain good performance in image analysis, in other cases, differences between deep learners and other classification method are very small. It may be

due to that we have not yet suitable network structure and we may not use suitable features for classification. The empirical fact that deep learners are good at recognition of images suggests that some kind of topological relations should be explicitly embedded into training data. So, here we propose a new feature selection method based on correspondence analysis, which calculates mapping attributes to points of multi-dimensional coordinates. The method can embed the topological relations between keywords and concepts into the data.

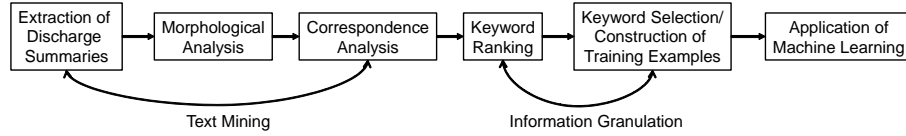


Fig. 1. Mining Process

3.3 Mining Process

Fig. 1 shows the proposed total mining process pipeline, composed of the following five steps. The first three methods can be called text mining process and the next two methods can be called information granulation process, which determines the granularity of knowledge for classification. Please note that these five methods are automated: after the users impose some conditions for text retrieval, such as collection period, and select the classification model, the process will start and the classification model will be generated.

Extraction of Discharge Summaries First, discharge summaries are extracted from the hospital information system as texts.

Morphological Analysis. Next, morphological analysis (MeCab [2]), is applied, which outputs a term matrix, that is, a contingency table for keywords and concepts are generated.

Correspondence Analysis. Thirdly, correspondence analysis is applied to a term matrix. Although high dimensional coordinates can be selected, since a very large table is obtained, we focus on two dimensional analysis which can be easily used for visualization. To each key word and concept, two dimensional coordinate is assigned.

Ranking. The coordinates of a concept and a keyword will be used to calculate the euclidean distance between them. We use the distance values for ranking of keywords to each concept: the smaller the distance is, the higher its ranking is.

Table 1. DPC top twenty diseases (fiscal year: 2015)

No DPC	Cases
1 Cataract (lateral)	445
2 Cataract (bilateral)	152
3 Type II Diabetes Mellitus (except for keto-acidosis)	145
4 Lung Cancer (with surgical operation)	131
5 Uterus Cancer (without surgical operation)	121
6 Lung Cancer (without surgical operation, chemotherapy)	
7 Uterus benign tumor	111
8 Lung Cancer (without surgical operation, with chemotherapy)	110
9 Shortage of Pregnancy	110
10 Injury of 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

Keyword Selection. We assume the number of keywords selected before the analysis, say 100. Usually, since the lower rank keywords give information about specific cases, the number of selected keywords will determine the granularity of induced knowledge. Then, keywords whose ranking are up to 100 are selected keywords for classification. Some keywords may overlap, so such overlapped keywords will be deleted. Then, training examples with a classification label and the value of selected keywords (binary attributes) are constructed.

Classification. Finally, classification learning methods will be applied. In this paper, we selected random forest [6], deep learning (multi-layer perceptron) (darch), Support Vector Machine (SVM) [4], Backpropagation Neural Network (BNN) [16] and decision tree (rpart) [13] for comparison.

Table 2. Number of Selected Keywords and 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

be removed, so only 1932 keywords were used for classification. Thus, some important keywords may be deleted due to the overlap if such words are frequently used at least in two diseases.

Classification. Finally, decision tree (package: rpart [13]), random forest (package: randomForest [6], SVM (kernlab [4]), BNN(package: nnet [16]) and Deep Learner (multi-layer perceptron) (darch⁴) were applied to the generated table. For parameters of Darch, the number of intermediate neurons are 10, (10,5), (40,10) and (100,10), whose epoch was 100. For all other packages, the default settings of parameters were used.

Evaluation Process. Evaluation process was based on repeated 2-fold cross validation [5]⁵.

First, a given dataset was randomly split into training examples and test samples half in half. Then, training examples was used for construction of a classifier, and the derived

⁴ Darch was removed from R package. Please check the githb: <https://github.com/maddin79/darch>

⁵ The reason why 2-fold is selected is that the estimator of 2-fold cross-validation will give the lowest estimate of parameters, such as accuracy and the estimation of bias will be minimized.

classifier was evaluated by using remaining test samples. The above procedures were repeated for 100 times in this experiment, and the averaged accuracy was calculated.

The number of keywords varied from 1 to 1000, selected according to the rank given by correspondence analysis. For analysis, two units of HP Proliant ML110 Gen9 (Xeon E5-2640 v3.2 2.6GHz 8Core, 64GBDRAM) was used.

Table 3. Experimental results (averaged accuracies)

#keywords	Darch	Darch	Darch	SVM	Rpart	Random	BNN
	one layer (20)	two layers (40,20)	two layers (80,20)			Forest	
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 3 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.

5 Discussion

5.1 Misclassified Cases

Fig 3 and 4 show confusion matrices of random forest and darch (multi-layer perceptron), where DPC codes are set in order which means that similar diseases are assigned to similar codes.⁶ Yellowed region indicates misclassified cases. It can be observed that whereas errors in random forest are located near the diagonal, errors in darch are more scattered. This suggests that random forest almost correctly classifies a case if similar dpc codes are grouped into one generalized class, while Darch has unexpected errors.

Prediction \ True	010000x 090000x Brain infarction	020110x x09700x Cataract (Un)	020110x x09700x Cataract (Bi)	020116x x09700x Retina Detachment (Un)	040040x x09000x Lung Cancer (with Surgical Operation)	040040x x09000x Lung Cancer (without Surgery)	040040x x09000x Lung Cancer (with Bronchoscopy)	040040x x09000x Lung Cancer (with Chemo Therapy)	04008x x09000x Pneumonia	060090x x09000x Liver Cancer (with Surgical Operation)	060340x x09000x Gallbladder Stone & Inflammation	070060x x09000x Severe Auto-immune Disease	10007x 000000x Type II DB	110280x x09000x Chronic Nephritis	120020x x09000x Uterus Cancer	12008x x09000x Uterus Benign Tumor	130180x x09000x Fetal Abnormalities	130230x x09000x Nonhodgkin	140010x x09000x Pregnancy (early)	16062x x09000x Injury of Knee or Elbow	Total	
010000x 090000x Brain infarction	17																				17	
020110x x09700x Cataract (Un)		20	9																			29
020110x x09700x Cataract (Bi)		33	14																			47
020116x x09700x Retina Detachment (Un)		1		28																		29
040040x x09000x Lung Cancer (with Surgical Operation)					8		1		1													11
040040x x09000x Lung Cancer (without Surgery)						16	10	3	2						2							33
040040x x09000x Lung Cancer (with Bronchoscopy)							1	16														17
040040x x09000x Lung Cancer (with Chemo Therapy)								2	28	1												32
04008x x09000x Pneumonia									17													17
060090x x09000x Liver Cancer (with Surgical Operation)									1	1												2
060340x x09000x Gallbladder Stone & Inflammation										11	4				1							16
070060x x09000x Severe Auto-immune Disease											1	8										9
10007x 000000x Type II DB													13									13
110280x x09000x Chronic Nephritis											1			15								16
120020x x09000x Uterus Cancer															1	4	1.4					6.4
12008x x09000x Uterus Benign Tumor																		4				4
130180x x09000x Fetal Abnormalities																					7	7
130230x x09000x Nonhodgkin																						1
140010x x09000x Pregnancy (early)																						21
16062x x09000x Injury of Knee or Elbow																						3
Total	20	65	39	29	6	18	30	36	24	13	15	14	21	19	6	4	7	21	3	0	390	

Fig. 3. Confusion Marix of Random Forest

⁶ DPC codes are three-level hierarchical system and each DPC code is defined as a tree. The first-level denotes the type of a disease, the second-level gives the primary selected therapy and the third-level shows the additional therapy. Thus, in the tables, characteristics of codes are used to represent similarities.

Prediction \ True	01006Dx 89903Dx	020110 x67700x 0	020110 x67700x 1	02016 x697A x60	040040x 87A00x	040040x 9900xx	040040 x6904x x	04004 x6904 00x	040080x 999400x	060050x 87A00x	060340 x0340x x	070560 x0990x x	10007 00000x x	110260x x69000x 0x	12002x x6994 0x	12006 x6902x 00x	120180x 01xxxx	130030x x9940x 0x	
	Brain infarction	Cataract (Uni)	Cataract (Bi)	Retina Detachment (Uni)	Lung Cancer (with Surgical Operation)	Lung Cancer (without Surgery)	Lung Cancer (with Bronchoscopy)	Lung Cancer (with Chemo therapy)	Pneu- monia	Liver Cancer (with Surgical Operation)	Gold- bladder Stone & Inflammation	Severe Auto- immune Disease	Type II DB	Chronic Nephritis	Uterus Cancer	Uterus Benign Tumor	Fetal Ab- normalities	Non- hodgekin	
01006Dx 89903Dx	13											1		3					
020110x x67700x		7	17																
020110x x67700x		25	41									1							
020160x x67700x		2		27															
040040x x67700x					5		1	1		1									
040040x x67700x						24	5	2						1					
040040x x67700x							3	10	1						2				1
040040x x67700x							3	2	26										1
040080x 999400x	2					2			11	2	1	7		3					
060050x 87A00x			1			3				4	8		1						1
060340x x0340x x	2	2	1	1					2	2	5								
070560x x09900x 0x	1							2					6	8					
100070x 000000x x														13	2				
110260x x69000x 0x		1	1			2			1	1			5	12					1
120020x x69940x 0x			1			1										3			
120060x x69020x 00x																5			
120180x 01xxxx									2							2			2
130030x x99400x 0x					1					1									20
142010x 199400x																			
160020x x01000x																			
Total	19	37	61	28	6	38	18	30	18	11	17	15	19	29	5	7	4	23	

Fig. 4. Confusion Matrix of Deep Learner

5.2 Execution Time

Tab. 4 shows an empirical comparison of repeated two-fold cross validations (100 trials). The times need for Random Forest and SVM were 183 and 156 minutes for 250 keywords, whereas Darch (20) required 672 minutes. For 1000 keywords, the times needed for Random Forest, SVM and Darch (20) were 261, 288, and 1101 minutes, 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. Although the performance of BNN and Deep Learners were high, the problem is that they need more time for computation.

5.3 Next Step

Experimental Validation shows that discharge summaries, as natural language texts, include enough information for classification of disease codes. And it also shows that 200 keywords for each disease are enough for classification. Thus, the next step is to apply automated knowledge acquisition process [14, 15] based on rough sets to the term matrix generated before application of learning methods

Table 4. 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.

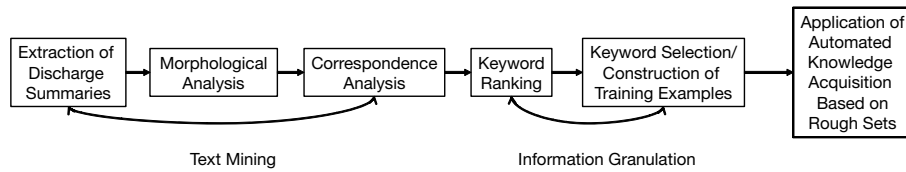


Fig. 5. Mining Process

6 Conclusion

Knowledge acquisition from medical experts is a very classical problem, but still a bottleneck for developing medical expert systems.

The reason why knowledge acquisition is difficult is that medical experts had difficulties in reformatting their knowledge into if-then rules. Thus, although deeper reasoning for experts emerged in the research of artificial intelligence in medicine in order to solve such bottle neck problems: it is still a difficult problem.

However, medical staff write down the process of clinical decision making as texts: it suggests that such knowledge about decision making can be represented as language texts, but not as rule-based. Thus, if we can analyze texts written by medical staff, we will be able to get knowledge about medical decision making.

We propose a method for construction of classifiers of disease codes from discharge summaries as follows. First, discharge summaries are extracted from hospital information system. Then, second, morphological analysis is applied to a set of summaries and a term matrix is generated. Third, correspond analysis is applied to the classification labels and the term matrix and generates two dimensional coordinates. By measuring the distances between categories and the assigned points, ranking of key words will be generated. Then, keywords are selected as attributes according to the rank, and training example for classifiers will be generated. Finally learning methods are applied to the training examples. Experimental validation was conducted by discharge summaries stored in Shimane University Hospital in fiscal year of 2015. The results shows that random forest achieved the best performance around 93% classification accuracy. The second best was deep learners with a small difference, around 91%. Thus, they suggests that discharge summaries includes enough information of clinical processes for each disease code. It will be our future work to acquire more structured knowledge from texts, such as diagnostic rules, rules for therapies, rules for prediction of outcome and so on.

References

1. Amisha, Malik, P., Pathania, M., Rathaur, V.: Overview of artificial intelligence in medicine. *Journal of family medicine and primary care* **8**(7), 2328–2331 (2019)
2. Ishida, M.: Rmecab. <http://rmecab.jp/wiki/index.php?RMeCabFunctions> (2016)
3. JONES, K.S.: A statistical interpretation of term specificity and its application in retrieval. *Journal of Documentation* **28**(1), 11–21 (1972). <https://doi.org/10.1108/eb026526>

4. Karatzoglou, A., Smola, A., Hornik, K., Zeileis, A.: kernlab – an S4 package for kernel methods in R. *Journal of Statistical Software* **11**(9), 1–20 (2004), <http://www.jstatsoft.org/v11/i09/>
5. Kim, J.H.: Estimating classification error rate: Repeated cross-validation, repeated hold-out and bootstrap. *Comput. Stat. Data Anal.* **53**(11), 3735–3745 (Sep 2009). <https://doi.org/10.1016/j.csda.2009.04.009>, <http://dx.doi.org/10.1016/j.csda.2009.04.009>
6. Liaw, A., Wiener, M.: Classification and regression by randomforest. *R News* **2**(3), 18–22 (2002), <http://CRAN.R-project.org/doc/Rnews/>
7. Luhn, H.P.: A statistical approach to mechanized encoding and searching of literary information. *IBM Journal of Research and Development* **1**(4), 309–317 (Oct 1957). <https://doi.org/10.1147/rd.14.0309>
8. Mares, M.A., Wang, S., Guo, Y.: Combining multiple feature selection methods and deep learning for high-dimensional data. *Transactions on Machine Learning and Data Mining* **9**, 27–45 (2016)
9. Nezhad, M.Z., Zhu, D., Li, X., Yang, K., Levy, P.: SAFS: A deep feature selection approach for precision medicine. *CoRR* **abs/1704.05960** (2017), <http://arxiv.org/abs/1704.05960>
10. Persidis, A., Persidis, A.: Medical expert systems: An overview. *Journal of Management in Medicine* **5**(3), 27–34 (1991). <https://doi.org/10.1108/EUM0000000001316>
11. Riaño, D., Wilk, S., ten Teije, A. (eds.): *Artificial Intelligence in Medicine - 17th Conference on Artificial Intelligence in Medicine, AIME 2019, Poznan, Poland, June 26-29, 2019, Proceedings, Lecture Notes in Computer Science*, vol. 11526. Springer (2019). <https://doi.org/10.1007/978-3-030-21642-9>
12. Shortliffe, E.: Medical expert systems—knowledge tools for physicians. *The Western journal of medicine* **145**(6), 830–839 (1986)
13. Therneau, T.M., Atkinson, E.J.: *An Introduction to Recursive Partitioning Using the RPART Routines* (2015), <https://cran.r-project.org/web/packages/rpart/vignettes/longintro.pdf>
14. Tsumoto, S.: Automated induction of medical expert system rules from clinical databases based on rough set theory. *Information Sciences* **112**, 67–84 (1998)
15. Tsumoto, S., Hirano, S.: Incremental induction of medical diagnostic rules based on incremental sampling scheme and subrule layers. *Fundam. Informaticae* **127**(1-4), 209–223 (2013). <https://doi.org/10.3233/FI-2013-905>, <https://doi.org/10.3233/FI-2013-905>
16. Venables, W.N., Ripley, B.D.: *Modern Applied Statistics with S*. Springer, New York, fourth edn. (2002), <http://www.stats.ox.ac.uk/pub/MASS4>, ISBN 0-387-95457-0