

Hybrid Approach to Data Mining Radiological Medical Records

William Claster, Subana Shanmuganathan and Nader Ghotbi
Ritsumeikan Asia Pacific University

wclaster@apu.ac.jp; s5subana@apu.ac.jp; nader@apu.ac.jp

Abstract

In this paper we extend previous work in an effort to extract meaningful information to distinguish necessary scans from that of what could be classified as undue exposure to radiation by analyzing medical records using text mining techniques. Here we develop a hybrid methodology for the analysis of medical records that consist of clinicians' notes (requesting for a scan) using statistical tools and Kohonen's self-organizing map (SOM) text mining techniques to look for heretofore invisible patterns to save patients from unwarranted scans. The medical data derive from patients' radiology department records where CT (Computed Tomography) scanning was used as part of a diagnostic exploration. The records are from the digital records of about 700 pediatric patients who underwent CT scanning (single and multiple) through a one-year period in 2004 at the Nagasaki University Medical Hospital in Japan. This approach led to a model based on SOM clusters and statistical analysis which allow for the prediction of when a particular medical screening procedure may be unnecessary. The procedure involves CT scans of patients. This is important because radiation at levels ordinarily used for CT scanning may pose significant health risks especially to children.

1. Introduction

Text-mining is applied in various fields to extract useful and previously unknown information contained in databases and texts. In the field of bioinformatics significant efforts are being made in genome sequencing, protein identification, medical imaging, and patient medical records. This study continues the efforts to mine patient medical records that consist of clinicians' notes in the form of free text. Harris et al. (2003)[1] developed a system to extract terms from clinical texts. Using natural language processing techniques, i.e., a parser, the MedLEE system (Friedman and Hripcsak, 1998)[2] turns free-text from patient records into an output with structured information. For example, it could identify patients with tuberculosis based on chest radiographs. It uses a corpus of controlled vocabulary developed from a collection of medical report. Chapman et al. (2004) use

text mining for automated fever detection from clinical records for biosurveillance. These as well as similar work discussed in [2], such as BIRADS UMLS, SNOMED, are useful in converting clinicians notes i.e., free text into some form of structured codes for medical diagnosis purposes. They use natural language parsers and a domain vocabulary (knowledge base) developed either using a corpus or stored expert knowledge. In this study we present our work in further investigations into potential approaches to distinguish medical terms that lead to unnecessary scans from clinicians' notes. These terms could be used to develop stepwise procedures to curtail unwanted scans and exposure to radiation especially in children

The increase in the use of medical radiation, especially in diagnostic CT scanning has raised many concerns over the possible adverse effects of procedures conducted in the absence of any serious risk/benefit analysis, especially where these procedures are carried out on children. Overuse can lead to unnecessary risk of exposure to radiation and may also contribute to rising health care costs[3][4][5].

Originally, prior to our investigation into the use of ANN based approaches, researchers at Nagasaki Hospital attempted to reevaluate the efficiency of CT scanning in both the diagnosis of acute appendicitis and also when used to detect possible injuries after acute head trauma using conventional methods[6]. As a result of that study a recommendation was made to the two departments studied. The recommendation was to employ guidelines and algorithms which present a stepwise set of clinical diagnostic methods and tools. The intention of this recommendation being that CT scans be reserved for patients that may be expected to benefit from them. However, in other departments, due to the lack of such a stepwise approach to diagnosis, many unnecessary CT scan have been and continue to be undertaken, and sound clinical judgment has been postponed until there is a confirmation by a CT scan. This was the initial impetus for our current work. The standard procedure adopted for requesting a scan at the Nagasaki Medical University Hospital as well as other medical practices is outlined in Fig. 1.

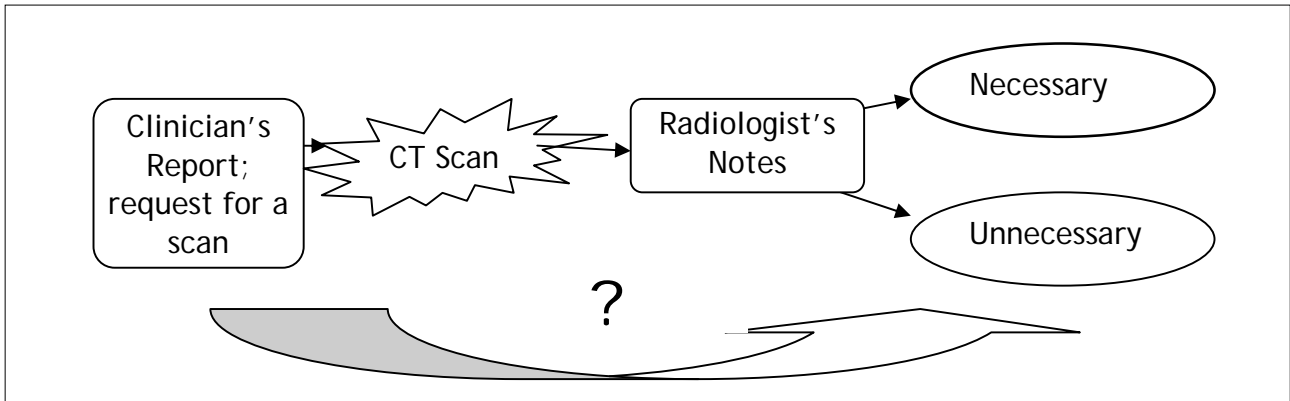


Fig. 1: Schematic diagram showing the standard procedure followed at Nagasaki Medical University Hospital and the expert classification on the necessity of a CT scan. We intend to develop tools to identify the features/ words that relate to unnecessary scans to curtail its overuse and thereby to overexposure to radiation, especially in children.

Much concern is focused on the lack of digitalized patient records[7] but in our study we were fortunate to have access to such records. With medical records obtained from Nagasaki Medical University Hospital Radiology Department's CT scanning database we preprocessed these records and emerged with a dictionary of 900 features (i.e., words) which were then used to search for clusters which could represent factors with predictive significance.

Originally, in our previous work (Claster, Nader, Shanmuganathan, 2007) we employed Kohonen Self Organizing Mapping to a sample of 50 of the free text medical records (clinicians' notes) out of a collection of 982 records obtained from the Nagasaki University Hospital. In the present study all 982 records were considered. In both the original study and the present study because of the free text nature of the data, the use of conventional analysis techniques became impracticable, as there were, as mentioned above, 900 words or features involved.

In that prior, 50 record study, we sought to look deeper into factors which may indicate unnecessary scanning. In this study, in addition to extending our work to the full 982 patient records we also develop a methodology for testing the analysis. We have used a k-fold validation design to test the proposed methodology. Finally we apply a novel approach to decision tree algorithms by focusing on the clusters discovered at the SOM layer of the analysis.

2.1 CT scan data and SOM based text mining

A thorough understanding of the indicators for the request to do a CT examination requires the analysis of huge amounts of text data in the medical records. Because of the unstructured form of these records, use of traditional statistical methods showed limited promise in

isolating factors that could accurately predict the patterns of CT scan usage. Unstructured text is a candidate for SOM text mining. SOM is a form of neural networks based on Kohonen's unsupervised learning algorithm. A commercial software package called Viscosity SOMine was employed to model the data. It provides a visualization tool that maps high dimensional inputs onto a two-dimensional map for easy visualization of the inputs that enhance detection of new knowledge in the form of patterns.

2.2 CT scan data preprocessing

Medical records were examined by breaking each patient record into its constituent words. We used a standard method of weighting the words which gives consideration to the frequency at which a word occurs in a document and also the overall frequency that the word occurs within the entire corpus. This method is known as tf-idf (see section 3.1). This allows us to recode text data as numerical data and thus makes it amenable to analysis with Kohonen mapping procedures. Our hypothesis is that information contained within the narrative text of medical records may determine whether a particular medical procedure (CT scans) would prove to be unnecessary.

Of the 1024 pediatric patient records extracted from the Nagasaki University Hospital, clinicians' notes and their outcome classified as either necessary or unnecessary scan by a medical expert from the radiologist comments were used in this SOM based clustering. The original clinicians' notes in Japanese were initially translated into English for this purpose. We then removed standard stop words (i.e. 'a', 'able', 'about', 'above' etc..) as well as common medical terms identified by a physician (i.e., 'abduction', 'advance', 'vessel') from the clinicians' notes. The 42 records that came out blank

through this process were removed altogether from this analysis. Consequently, a matrix of word x record no. was created in which the rows consisted of all the words in the text corpus of patient records, to apply the tf x idf formula discussed in the methodology section. For further details on the process and formula see Cluster et al (2007)[8]. From this matrix, words that were found to be useless in the analysis were again removed and a new matrix of word x patient record numbers was created. The records that lost all the words in the process were labeled as ‘no comments’. The records that had radiologist notes unclassifiable either as necessary or unnecessary by the medical expert were also tagged with ‘no comments’. Using the final matrix we created a 2000 node SOM (Figs 2 and 5). The validation performed on the SOM clustering is discussed in the next section.

3. Methodology: Cluster Exploration, Model Development and Testing

An expert in the medical field (in this case a physician) indicated whether each particular scan was necessary or unnecessary and thus we had a classification for the data. Viscovery SOM discovered clusterings on the dataset ranging from 2 clusters to more than 30. (see Fig 5). When we explored the 2-cluster SOM (Fig2) we were able to identify a cluster of records that were 98% necessary (C1) and a cluster of records that were 97% unnecessary (C2) -where necessary and unnecessary refer to whether a CT scan was deemed necessary or not by the expert. The two cluster profiles are shown in Fig 4

We analyzed each cluster and developed a methodology (described below) to weight a word in the text according to how often it occurred in either cluster of the 2 cluster SOM (Fig 2). (note: this does not refer to tf-idf but a weighting used for another purpose described below). In short this can be described as: a word appearing only in the cluster C1 was given a high positive weight and a word appearing only in the cluster (C2) a high negative weight. Then the word weights for all the words contained in a particular record were summed to determine a record classification value (“rc”) for that record. This record classification value was then used to predict whether a record’s classification belonged to the class of necessary scans or to the class of unnecessary scans. A K-fold cross-validation procedure was later employed as a means of model verification.

		Predicted Classification	
		necessary	unnecessary
Actual Classification	necessary	506	105
	unnecessary	108	253

Table 1. Averaged results of K-fold cross validation

3.1 Model

The tf-idf weight (term frequency-inverse document frequency) is a statistical measure used to evaluate how important a word is to a document in a collection or corpus. The term frequency is given by:

$$tf_i = \frac{n_i}{\sum_k n_k}$$

with n_i being the number of occurrences of the considered term, and the denominator is the number of occurrences of all terms. The inverse document frequency is given by:

$$idf_i = \log \frac{|D|}{|\{d : d \ni t_i\}|}$$

with $|D|$ the total number of documents in the corpus and

$|\{d : d \ni t_i\}|$: the number of documents where the term

t_i appears (that is $n_i \neq 0$). Then

$$tfidf = tf \cdot idf .$$

Now define a word prediction vector for word w_i by defining j th component of the word prediction vector (“wpv_{ij}”) for word w_i in cluster j as:

$$wpv_{ij} = \frac{\sum_{\text{all records in cluster } j} (tf - idf)}{n_{i,j}}$$

In other words, wpv_{ij} is the mean of the tf-idf over all the records in the j th cluster for word w_i .

Next letting K_i be the cluster with the maximum wpv_{ij} for word w_i (over all j clusters), define the word prediction weight (“wpw_i”) for word w_i over all clusters to be:

$$wpw_i = wpv_{i,K_i} - \sum_{j \neq K_i} wpv_{ij}$$

(note, that in our case there are just two clusters 1 and 2, and therefore this becomes just wpv_{i1} - wpv_{i2} which will be either positive or negative).

These word prediction weights, wpw_i, are used to establish an overall value for any record. This gives us a way of taking a new document and assigning a value (and therefore a classification as a necessary or unnecessary CT scan) based on the following scheme.

Define the record value v as

$$v = \sum_{\text{words in the record}} wpw_i$$

Then define rc to classify a record as necessary or unnecessary by

$$rc = \begin{cases} \text{necessary} & \text{if } v \geq 0 \\ \text{unnecessary} & \text{if } v < 0 \end{cases}$$

3.2 Model Testing

The above classification model was tested in a K-fold cross-validation procedure. We subdivided the data into 3 subsets and the cross validation process was carried out 3 times, after which the results of the folds were averaged.

4. Results

In previous work we were able to identify through the methods of text mining explained earlier, a series of keywords within the CT scan referral rationale. The statistical strength assigned to the keywords led to their separation into three sets which had a strong association with a positive finding by radiologists, a strong association with a negative finding by radiologists, or a weak association with both a positive and a negative finding (Fig 4).

4.1 Classifying and Testing

We conducted a 3-fold cross validation procedure and arrived at Table 1. This table shows a false-positive error rate of 11% and a false-negative error rate of 11% and an overall error rate of 22%. Although this accuracy is substantial we may improve upon it by the elimination of words which occur in higher frequencies in both clusters. Modification of the wpw_{ij} weight assignment may also contribute to a reduction in either false positives and/or false negatives.

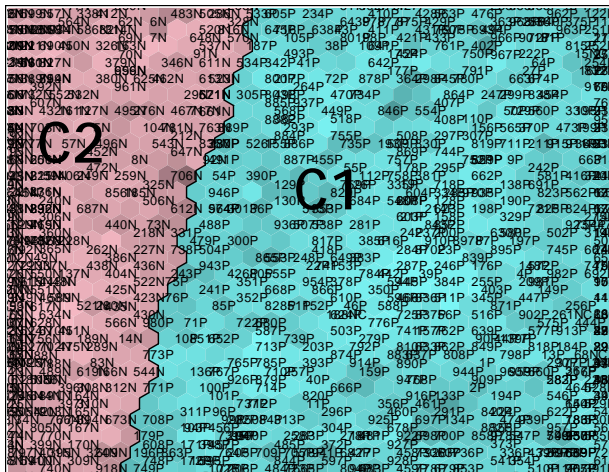


Fig 2; SOM of physician's CT referral rationale (text being mined) segregated into positive versus negative

outcome clusters. Additional preference was given to have the positive and negative notes clustered together. Cluster C1 contains 98%P (necessary scans) and Cluster C2 contains 95% N (unnecessary scans).

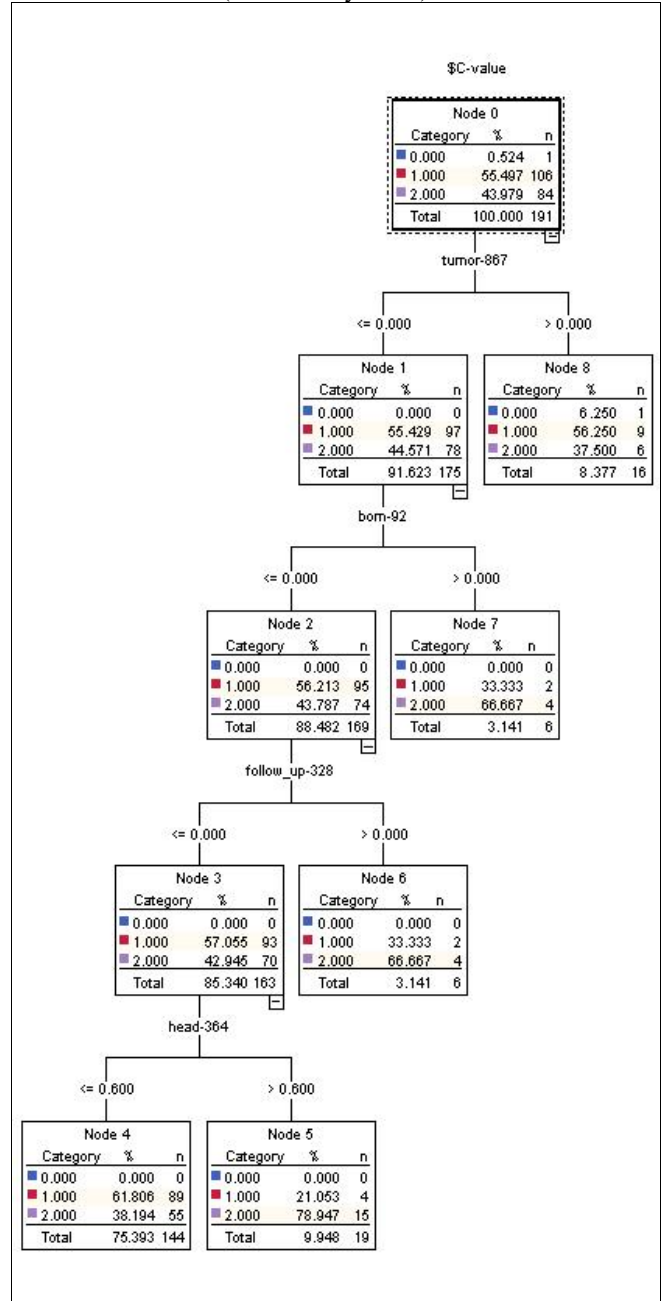
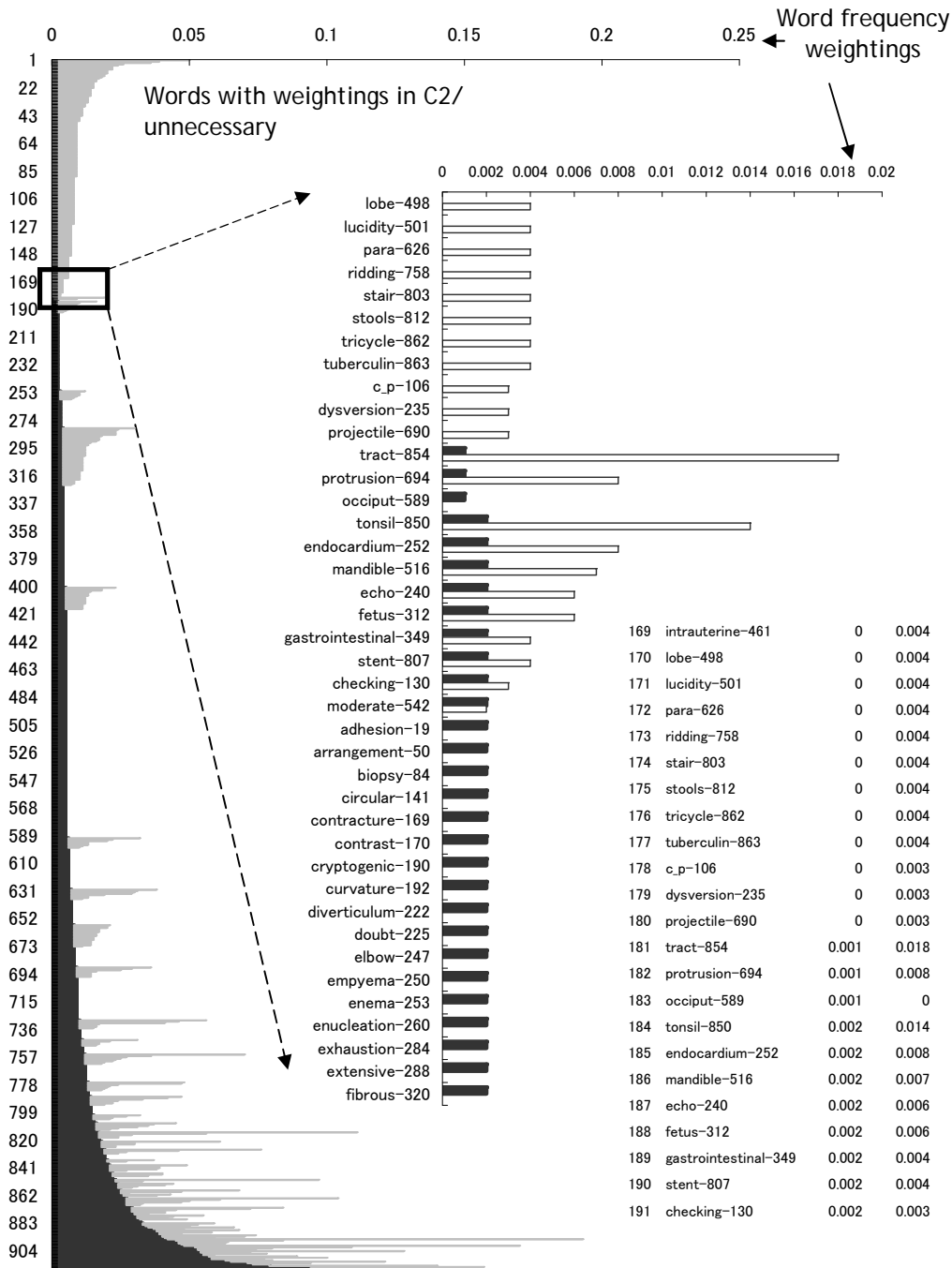


Fig 3; C5.0 chart for cluster C1 (expanded view). For example, node 2 indicates that when tumor ≤ 0 then 55.4% of the records were necessary and 44.6% were unnecessary.



The grey areas represent mean word frequency weightings in C2 unnecessary scan. The black areas represent those of C1 necessary scan, according the areas with both represent those weightings present in both C1 and C2.

Fig. 4. C 1 (necessary) word weights (mean, minimum, maximum and sum). C 2 (unnecessary) word weights (mean, minimum, maximum and sum)

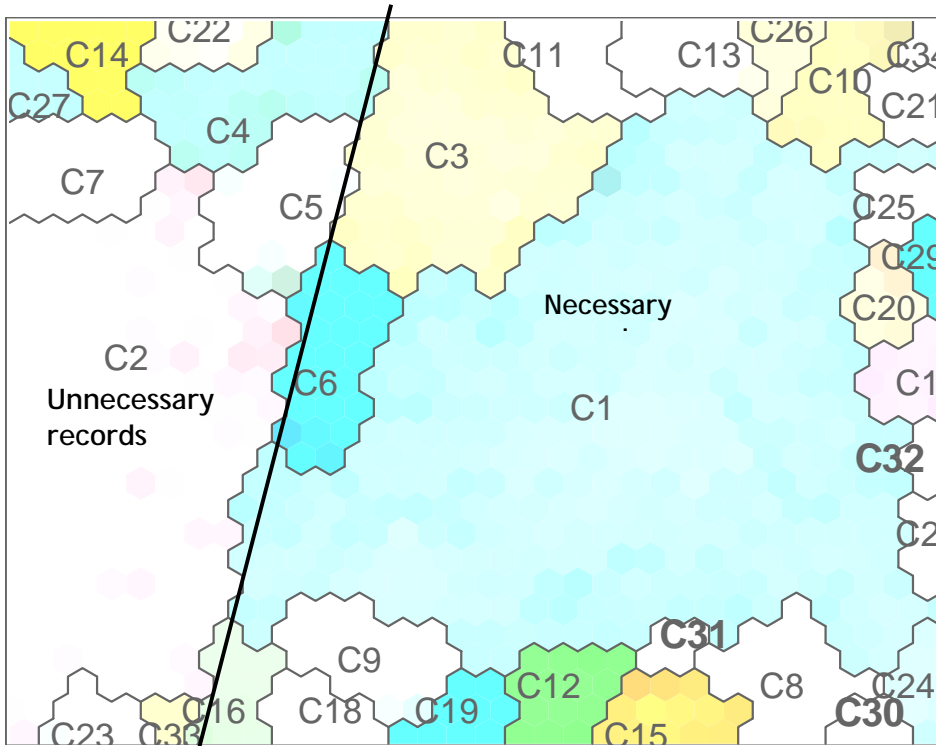


Fig 5. Positive records further subdivided into 25 clusters and negative records subdivided into 9 clusters.

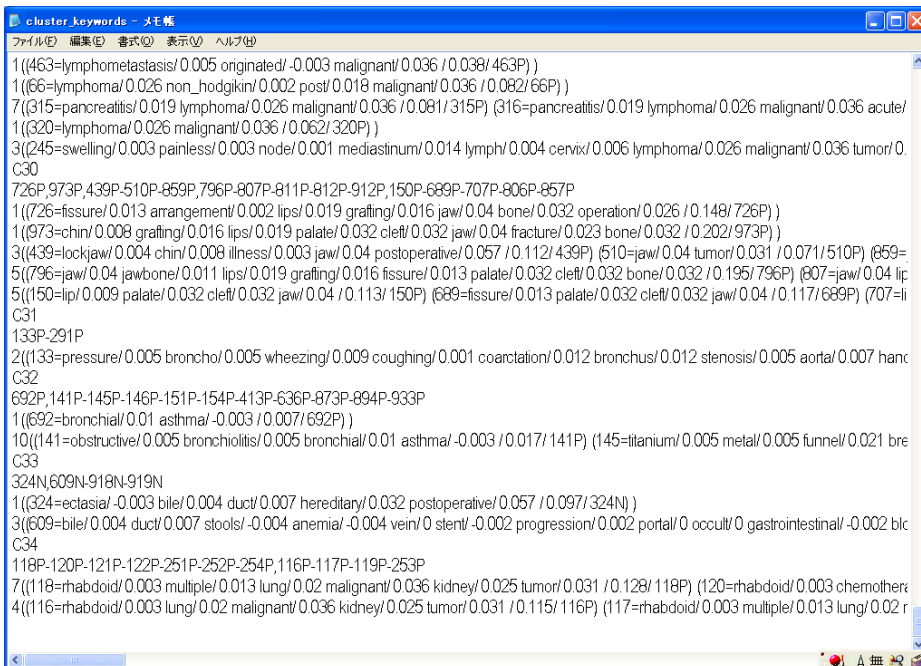


Fig. 6. Constituent words (including record number, word, weights, and necessary/ unnecessary classification) within clusters 30 to 34.

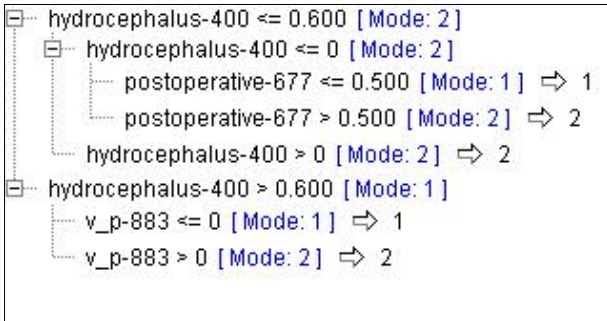


Fig. 7. C5.0 chart for subgroup cluster number 30. 1 indicates the predication of a necessary scan and 2 indicates the prediction of an unnecessary scan.

4.2 Clusters

We investigated the subgroups within C1 (necessary) and C2 (unnecessary) clusters to see whether these subgroups could be developed into a C5.0 charts. Based on Viscovery's, (commercial software) clustering suggestions, C1 and C2 clusters were further divided into 25 and 9 subgroups respectively (Fig 5) and then their word groupings (Fig 6) were analyzed for any possible scenarios of developing a C5.0 chart. Most of the groupings appeared to consist of a uniform and unique set words relating to a particular type of disease (ENT, pulmonary) accident (involving a vehicle, fall from a tree, etc) or birth defects. For example, clusters 30 and 31 of C1 have words relating to neural disorders. See Figure 5. Similarly, cluster 32 words relate to orthopedic (lower facial).

We ran a C5.0 algorithm to produce a decision tree (Figs 3 and 7). C5.0 is a commercial classification algorithm used to generate a decision tree using the idea of information entropy. It is not restricted to producing binary trees. It is hoped that a tree could provide feedback to a medical practitioner that may be included as one factor in the decision to request a CT scan and we will explore this in future work.

4.3 Conclusions

This shows that medical doctors may be able to consider some pre-test factors as predicted strength of the test, before requesting a CT scan for children. Our analysis gives a preliminary methodology for predictive record classification. The 11% false-negative rate is large but there may be strategies for pruning this error rate. Decision charts based on SOM based clustering can be employed for more focused results.

5. Future work

Further research is underway to study an additional 10,000 records. With these records we hope to explore other weighting systems and by including time series analysis to develop an expanded hybrid methodology to include seasonal effects in order to achieve improved accuracy. Using the same data, we will compare the current methodology with a neural network classification scheme as well as modifying the SOM clustering to a K-means clustering algorithm. Additional research is underway to measure the effectiveness of SOM based decision charts. We believe it is possible to design a form of text mining system that helps with such decision making when a medical doctor is considering whether or not a CT scan may be helpful in reaching a diagnosis. This text mining system can be fed with the hospital's own data so that patterns of association between clinical information and radiological findings are determined, and help with decision-making further on.

6. Acknowledgments

The authors wish to express their sincere gratitude to Professor Monte Cassim who has been a tremendous inspiration for all of us here at Discovery Laboratory, especially, into text mining clinical data. Nagasaki University Hospital clinical staff members are acknowledged for permission to use their data in this study

7. References

- [1] Harris, M. R., Sovova, G.K., Johnson, T.M., and Chute, C.G. (2003). A Term Extraction Tool for Expanding Content in Domain of Functioning, Disability, and Health: Proof Tool for Expanding Content in the Domain of Functioning, Disability, and Health: Proof of Concept," *Journal of Biomedical Informatics*, 36, 250-259.
- [2] Identification of suspected tuberculosis patients based on natural language processing of chest radiograph reports *Report ID: DBMI-1996-038 Authors: Jain NL, Knirsch CA, Friedman C, Hripcsak G*
- [3] Frush DP, Donnelly LF, Rosen NS. Computed tomography and radiation risks: what pediatric health care providers should know. *Pediatrics*, 112: 951-957, 2003.
- [4] Brenner DJ, Ellison CD, Hall EJ, Berdon WE. Estimated risks of radiation induced fatal cancer from pediatric CT. *AJR*, 176: 289-296, 2001.

-
- ^[5] Roebuck DJ. Risk and benefit in pediatric radiology. *Pediatr Radiol*, 29: 637-640, 1999.
- ^[6] Ghotbi N, Morishita M, Ohtsuru A and Yamashita S, Evidence-based Guidelines Needed on the Use of CT Scanning in Japan in Japan Medical Association Journal (JMAJ) Vol. 48, No. 9 September 2005
- ^[7] Walsh S H. The Clinician's Perspective on Electronic Health Records. *BMJ* 2004; 328;1184-1187.
doi:10.1136/bmj.328.7449.1184
- [8] William Claster, Subana Shanmuganathan, and Nader Ghotbi 2007, Text Mining in Radiological Data Records: An Unsupervised Neural Network Approach in proceedings of the First Asia International Conference on Modelling & Simulation (AMS2007 Asia Modelling Symposium 2007) held in conjunction with Thailand's 11th Annual National Symposium on Computational Science and Engineering (ANSCSE-11), published by the IEEE Computer Society (order no., P2845) ISBN 0-7695-2845-7 Eds., David Al-Dabass, Richard Zobel, Ajith Abraham and Steve Turner. Prince of Songkla University, Phuket, Thailand. 27-30 March 2007 pp329-333