Chapter 2
The Use of Artificial Intelligence Systems for Support of Medical Decision-Making

William Claster
Ritsumeikan Asia Pacific University, Japan

Nader Ghotbi
Ritsumeikan Asia Pacific University, Japan

Subana Shanmuganathan
Auckland University of Technology, New Zealand

ABSTRACT

There is a treasure trove of hidden information in the textual and narrative data of medical records that can be deciphered by text-mining techniques. The information provided by these methods can provide a basis for medical artificial intelligence and help support or improve clinical decision making by medical doctors. In this paper we extend previous work in an effort to extract meaningful information from free text medical records. We discuss a methodology for the analysis of medical records using some statistical analysis and the Kohonen Self-Organizing Map (SOM). The medical data derive from about 700 pediatric patients' radiology department records where CT (Computed Tomography) scanning was used as part of a diagnostic exploration. The patients underwent CT scanning (single and multiple) throughout a one-year period in 2004 at the Nagasaki University Medical Hospital. Our approach led to a model based on SOM clusters and statistical analysis which may suggest a strategy for limiting CT scan requests. This is important because radiation at levels ordinarily used for CT scanning may pose significant health risks especially to children.

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 clinician notes in the form of free

DOI: 10.4018/978-1-60566-266-4.ch002
text. Harris et al (2003) developed a system to extract terms from clinical texts. Using natural language processing techniques, i.e., a parser, the MedLEE system, Jain et al (1996) turned free-text from patient records into an output with structured information. For example, this system may identify patients with tuberculosis based on chest radiographs. To do this it uses a corpus of controlled vocabulary developed from a collection of medical reports. This, as well as similar work discussed below such as BIRADS UMLS, SNOMED, are useful in converting clinician notes as 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.

The gold standard in text mining is natural language processing, which aims to include semantic information in the text mining task. The full realization of this goal is still on the distant horizon however serious efforts have already achieved some success. These include AQUA, PROTEUS-BIO, and SemRep. AQUA (A Query Analyzer) is an underspecified semantic interpreter that was originally formulated for processing MEDLINE queries. PROTEUS-BIO applies to web documents on infectious disease outbreaks. It mines semantic predications relevant to this domain and stores them in a database. The database is then available to users. SemRep is being developed to recover semantic propositions from biomedical research literature. It again focuses on MEDLINE citations. SemRep utilizes underspecified syntactic analysis and structured domain knowledge. In addition to investigations that consider semantics, non semantic approaches are yielding significant progress as well. Vector space models, neural networks, kernel methods, decision trees and rule induction, and probabilistic models are all being used for classification without strong emphasis on semantic characteristics and are yielding promising and interesting results as applied to text mining.

**BACKGROUND**

The advent of computed tomography (CT) has revolutionized diagnostic radiology (Figure 1). Since the inception of CT in the 1970s, its use has increased rapidly. It is estimated that more than 62 million CT scans per year are currently obtained in the United States, including at least 4 million for children (Brenner & Hall, 2007). 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 (UNSCEAR, 2000). According to a survey conducted in 1996 (White, 1996) the number of CT scanners per 1 million population was 26 in the United States and 64 in Japan. The growth of CT use in children has been driven primarily by the decrease in the time needed to perform a scan — now less than 1 second — largely eliminating the need for anesthesia to prevent the child from moving during image acquisition (Frush et al, 2003). Overuse of diagnostic CT radiation, which deliver radiation doses 50 to 200 times higher than most X-rays, can lead to an increased risk of cancer. Additionally, it may lead to an unnecessary rise in health care costs (Roebuck, 1999; Ghothbi et al, 2005; Walsh, 2004).

In this system a motorized table moves the patient through the CT imaging system. At the same time, a source of x-rays rotates within the circular opening, and a set of x-ray detectors rotates in synchrony on the far side of the patient. The x-ray source produces a narrow, fan-shaped beam, with widths ranging from 1 to 20 mm. In axial CT, which is commonly used for head scans, the table is stationary during a rotation, after which it is moved along for the next slice. In helical CT, which is commonly used for body scans, the table moves continuously as the x-ray source and detectors rotate, producing a spiral or helical scan. The illustration shows a single row of detectors, but current machines typically have multiple rows of
The Use of Artificial Intelligence Systems for Support of Medical Decision-Making

Figure 1. The basics of CT

Figure 1. The basics of CT

detectors operating side by side, so that many slices (currently up to 64) can be imaged simultaneously, reducing the overall scanning time. All the data are processed by computer to produce a series of image slices representing a three-dimensional view of the target organ or body region (White, 1996).

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 (Deboeck & Kohonen, 1998a). 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 is generally postponed until viewing the results of 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 and classification by field expert is outlined in Figure 2.

In the past much concern has focused on the lack of digitalized patient records, but with the digital revolution in full sway this is becoming a non-issue (Walsh, 2004). Using 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.

In our previous work (Claster et al, 2007) we employed Kohonen Self Organizing Mapping to analyze a sample of 50 of the free text medical records (clinician 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 900 words or features involved. In the prior 50 record study we sought to look deeper into factors which might indicate
The Use of Artificial Intelligence Systems for Support of Medical Decision-Making

Figure 2. Schematic diagram showing the standard procedure followed at Nagasaki Medical University Hospital and the expert classification on the necessity of a CT scan.

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. A SOM is a feed forward neural network (Figure 3), which uses an unsupervised training algorithm to perform non-linear regression. Through a process called self-organization the network configures the output data into a display of topological representation, where similar input data are clustered near each other. At the end of the training SOM enables analysts to view any novel relationships, patterns or structures in the input vectors. The topology preserving mapping nature of SOM algorithm is highly useful in projecting multi-dimensional data sets into low dimensional displays, generally into one- or two-dimensional planes. Thus SOMs can be used for clustering as well as visualization of multi-dimensional data sets (Deboeck & Kohonen, 1998b).

The SOM techniques are successfully applied to visualize and cluster large volumes of complex statistical data sets of many real world problems such as pattern recognition, image analysis, process monitoring and control, and fault recognition. As SOM methods are based on an unsupervised training algorithm, they could be used for data clustering without knowing the class membership of the input data (Simula & Vesanto, 1999). Traditional methods (i.e. simple statistical methods) that are useful in summarizing low-dimensional data sets (mean value, smallest and highest values), are seen to be less effective in visualizing multi-dimensional (i.e. multivariate) data sets (Deboeck, 1998; Deboeck & Kohonen, 1998b).

The self-organizing map (SOM) algorithm, first introduced by Tuedo Kohonen in 1982 was developed from basic modeling information of the human brain’s cortical cells, as they were known from the neuro-physiological experiments of the late twentieth century. The processing of synaptic connections between the cortex cells in the human brain is based upon the nature of the sensorial stimuli; hence different patterns of sensorial signals converge at different areas within the brain’s cortex cells. Because of this different individual neurons or groups of neurons become sensitive
to different sensorial stimuli, and neighboring neurons also learn to respond to similar patterns of signals (i.e., visual, auditory, somatosensory, etc). Despite this realization our knowledge on the associative areas of signals and the other different tasks involved with the rest of the cortical area is relatively poor. Only ten percent of the total cortical area is described as involved with primary sensorial signals. The planning of actions is assumed to take place in the frontal lobe.

Kohonen's SOM applications to real world problems range across many disciplines, mainly in the field of knowledge discovery. SOM ability to discover implicit knowledge from numerical data displaying the input vectors on low dimensional grid structures is significant. The following are some of the major identified areas of SOM applications (Deboeck & Kohonen, 1998):

1. Classification, clustering, and/or data reduction;
2. Visualization of the data;
3. Decision-support;
4. Hypothesis testing;
5. Monitoring system performance;
6. Lookup of (missing) values;
7. Forecasting.

Traditional statistical methods consist of limited abilities for revealing structures, relationships and novel patterns in low dimensional data sets. Two to three dimensional data sets can be visualized by using simple two- to three-dimensional graphs. But with multi dimensional data sets, plotting a vector or analyzing the relationships between different vectors by simple graphs is not possible. Thus other methods are needed to visualize such multi dimensional data sets.

In existing data visualization methods, the different components contributed by each and every dimension are integrated into the one final result. The major drawback experienced with conventional methods is that they are unable to reduce the amount of data within large sets, as processing becomes incomprehensible. However, they can be used to display simple summaries of data sets (Deboeck & Kohonen, 1998c). Data clustering is one such operation through which similar data items are categorized or grouped together and it is one of the best possible methods available for reducing large volumes of data for visualization purposes. Clustering is described as similar to information processing in humans and preferred over projection methods. The goal of projection methods is to represent the input data in a chosen low dimensional space, where certain properties of the structure of original data are preserved as faithfully as possible to the original values. Thus these projections can be used to visualize a high dimensional data set if a sufficiently enough low dimensionality is chosen for output display. Clustering can be automated to classify different categories and automation also reduces bias and errors in the grouping process. Traditional clustering methods can be classified into two basic types: hierarchical and (ii) non-hierarchical. On the other
hand, projection methods can also be classified into two basic types: linear and (ii) non-linear.

In our example a commercial software package called Viscovery SOMine was employed to model the data. This 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.

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 below). 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, clinician 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 clinician 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). 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 (Figures 4 and 7). The validation performed on the SOM clustering is discussed in the next section.

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. When we explored the 2-cluster SOM (Figure 4) 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 Figure 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 (Figure 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
The U se of Artificial Intelligence Systems for Support of Medical Decision-Making

Figure 4. SOM of physician's CT referral rationale (text being mined) segregated into positive versus negative outcome clusters

then used to predict whether a particular record 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 (Table 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:

\[ idf_i = \log \left( \frac{|D|}{|\{d : d \ni t_i\}|} \right) \]

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 \( tfidf = tf \times idf \).

Now define a word prediction vector for word \( w_i \) by defining jth component of the word prediction vector (“wpv\(_j\)”) for word \( w_i \) in cluster \( j \) as:

\[ wpv_{ij} = \sum_{\text{all records in cluster } j} \frac{(t_{ij} - idf_j)}{n_{ij}} \]

In other words, wpv\(_{ij}\) is the mean of the tf-idf over all the records in the jth cluster for word \( w_i \).

Next letting \( K_i \) be the cluster with the maximum wpv\(_i\) for word \( w_i \) (over all j clusters), define the word prediction weight (“wpw\(_i\)”) for word \( w_i \) over all clusters to be:

\[ wpw_i = wpv_{ik_i} - \sum_{j \neq k_i} wpv_{ij} \]

(note, that in our case there are just two clusters 1 and 2, and therefore this becomes just wpv\(_w \) - wpv\(_v\), which will be either positive or negative).
The Use of Artificial Intelligence Systems for Support of Medical Decision-Making

Figure 5. 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.

These word prediction weights, wpw, 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.
The Use of Artificial Intelligence Systems for Support of Medical Decision-Making

Table 1. Averaged results of K-fold cross validation

<table>
<thead>
<tr>
<th>Actual Classification</th>
<th>Predicted Classification</th>
</tr>
</thead>
<tbody>
<tr>
<td>necessary</td>
<td>506</td>
</tr>
<tr>
<td>unnecessary</td>
<td>105</td>
</tr>
<tr>
<td>necessary</td>
<td>108</td>
</tr>
<tr>
<td>unnecessary</td>
<td>253</td>
</tr>
</tbody>
</table>

Define the record value $v$ as:

$$v = \sum_{w \in \text{the record}} wp_w,$$

Then define $r_c$ to classify a record as necessary or unnecessary by

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

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.

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 (Figure 6).

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 $wp_v$ weight assignment may also contribute to a reduction in either false positives and/or false negatives.

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).

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 SOM clustering suggestions (generated with Viscovery commercial software) C1 and C2 clusters were further divided into 25 and 9 subgroups respectively (Figure 7) and then their word groupings (Figure 8) 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 8). Similarly, cluster 32 words relate to orthopedic (lower facial).
The Use of Artificial Intelligence Systems for Support of Medical Decision-Making

Figure 6. C1 (necessary) word weights (mean, minimum, maximum and sum). C2 (unnecessary) word weights (mean, minimum, maximum and sum)

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

We ran a C5.0 algorithm to produce a decision tree (Figures 5 and 9). 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.
The Use of Artificial Intelligence Systems for Support of Medical Decision-Making

Figure 7. Positive records further subdivided into 25 clusters and negative records subdivided into 9 clusters

CONCLUSION

This discussion 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 are strategies for pruning this error rate.

Figure 8. Constituent words (including record number, word, weights, and necessary/unnecessary classification) within clusters 30 to 34
The Use of Artificial Intelligence Systems for Support of Medical Decision-Making

Figure 9. C5.0 chart for subgroup cluster number 30. 1 indicates the prediction of a necessary scan and 2 indicates the prediction of an unnecessary scan.

- hydrocephalus-400 <= 0.600 \( \rightarrow \) \( [\text{Mode: 2}] \)
- hydrocephalus-400 <= 0 \( \rightarrow \) \( [\text{Mode: 2}] \)
  postoperative-677 <= 0.500 \( \rightarrow \) \( [\text{Mode: 1}] \) \( \rightarrow \) \( 1 \)
  postoperative-677 > 0.500 \( \rightarrow \) \( [\text{Mode: 2}] \) \( \rightarrow \) \( 2 \)
  hydrocephalus-400 > 0 \( \rightarrow \) \( [\text{Mode: 2}] \) \( \rightarrow \) \( 2 \)
- hydrocephalus-400 > 0.600 \( \rightarrow \) \( [\text{Mode: 1}] \)
  v_p-883 <= 0 \( \rightarrow \) \( [\text{Mode: 1}] \) \( \rightarrow \) \( 1 \)
  v_p-883 > 0 \( \rightarrow \) \( [\text{Mode: 2}] \) \( \rightarrow \) \( 2 \)

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.

ACKNOWLEDGMENT

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.

REFERENCES


