SOM – Feature Extraction from Patient Discharge Summaries

D. J. Tufts-Conrad[1], A. N. Zincir-Heywood[1], D. Zitner[2]

[1]Dalhousie University, Faculty of Computer Science, Halifax, NS, Canada
[2]Dalhousie University, Department of Medical Informatics, Halifax, NS Canada

dyan@cs.dal.ca, zincir@cs.dal.ca, david.zitner@dal.ca

Keywords: Neural networks for medical informatics, computer application for medical informatics, health care information systems

Contact Address:
Dr. N. Zincir-Heywood
Dalhousie University
Faculty of Computer Science
Halifax, NS B3H 1W5
Canada

Phone: +1 (902) 494-3157
Fax: +1 (902) 492-1517
E-mail: zincir@cs.dal.ca

SOM – Feature Extraction from Patient Discharge Summaries

Abstract

In each Canadian province, hospitals collect information, at discharge, on the hospital stay of each patient. The information is collected in the form of a patient discharge abstract (PDA) and sent to the Canadian Institute for Health Information. The PDA uses the ICD-10-CA code standard to outline the assigned diagnoses for the patient's condition and the procedures that were performed. One compulsory piece of information in the PDA is the identification of the "most responsible diagnosis" (MRDx) – that diagnosis considered to be the most significant condition of the patient that caused the greatest length of stay in hospital.

This research investigates the potential for automating the process of feature extraction from a narrative patient discharge summary (PDS) to support the classification of the MRDx for a PDA. Unsupervised neural networks – Self-Organizing Maps (SOM) – are effective for classification tasks based on noisy input patterns. Here a hierarchical architecture of SOMs is used to identify semantic similarities encoded in the original information and visualize the characteristics of an MRDx.

I. Introduction

During the time a patient is in hospital an extensive amount of data is collected for the hospital record. This includes ongoing progress notes, laboratory notes, results from any tests run including X-rays and electrocardiograms and operating reports for any surgeries. Once a patient has been discharged from the hospital, a Patient Discharge Summary (PDS) will be dictated by the attending practitioner, which summarizes the patient's stay in the hospital. Information recorded in the PDS includes the admission diagnosis, discharge diagnosis, course in hospital, medications on discharge and recommended follow-up. It is dictated in a narrative manner by the practitioner, hence the PDS is also a form of communication between the practitioner, who looks after the patient while in hospital, and the family physician.

Each hospital record of a discharged patient is subject to a detailed abstraction process by a health records reviewer. The reviewer collects information on the diagnosis and manually abstracts the chart for use by the Canadian Institute for Health Information (CIHI). These medical diagnoses are coded by reviewers according to ICD-10-CA codes[1]. The ICD-10-CA classifies diseases, injuries, causes of death, and external causes of injury and poisoning. Classifications use an alphanumeric scheme and categorize the codes in 23 chapters. These chapters correspond primarily to the body's physiological systems, such as the digestive, respiratory and circulatory system.

[1] In May of 1999, CIHI developed the Canadian version of the ICD-10-CA standard, known as the *International Statistical Classification of Diseases and Health Problems, Tenth Revision, Canada*. On April 1, 2001, the ICD-10-CA supplanted the use of the ICD-9-CM.

Moreover, large hospitals will employ 20 or more individuals to abstract the charts of discharged patients [3]. By 1999 the Queen Elizabeth II Health Sciences Centre spent more than $1.3 million dollars annually for chart abstraction to support the CIHI discharge abstract [9]. The interrater reliability for the PDA's is currently not known. Thus, the use of automated support systems for coding may provide an opportunity for substantial cost savings and improved reliability.

On the other hand, formation of such automatic systems is not as straightforward as it seems. Many individuals in a hospital are responsible for the creation of a PDS introducing variation into these narrative communications; additional variation in the scope and extent of these documents adds even further complexity to the creation of standard features to represent the document using standard language processing techniques. Moreover, by examining the PDSs and PDAs, it can be seen that there are many diagnosis types identified on a PDA for a given PDS where these may have zero, one, or multiple codes assigned.

At a minimum, each PDA has one Most Responsible Diagnosis (MRDx) and one Admission Diagnosis assigned for each PDS. These two diagnosis types share a one-to-one relationship with the discharge summary. This work focuses on providing automatic support in order to determine a patient's *Most Responsible Diagnosis* (MRDx), defined as that condition which kept the patient in hospital the longest. To this end, the feasibility of using the unsupervised neural network technique of Self Organizing Feature Maps (SOM) is explored to develop an automatic feature extraction system that uses PDSs for the formation of the MRDx in PDAs.

The remainder of this paper is organized as follows. Section 2 introduces the SOM, unsupervised learning algorithm, used for feature extraction. Section 3 describes the architectural overview, whereas experimental settings and results are detailed in section 4. Finally, the conclusions are drawn in section 5.

II. Feature Extraction

As indicated above, in this research, the aim is to investigate the potential for automating the process of feature extraction from a Patient Discharge Summary (PDS) in order to support automatic coding of the MRDx in patient discharge abstracts.

In order to achieve this, the first step is the identification of an encoding for the original information (MRDx) such that pertinent features may be decoded most efficiently. This information is then used to measure the similarity between the characteristics of any given PDS to an MRDx. Thus, the core of the approach is to automate the identification of typical MRDx characteristics. To this end, an unsupervised learning system – Self Organizing Feature Map (SOM) – is employed to detect and visualize the characteristics of an MRDx.

The SOM algorithm is a vector quantization algorithm that posses several properties that make it very convenient for expressing relationships between different groups of data. For example: an efficient update scheme and ability to express topological relationships between similar properties of the data. These properties have made the SOM a popular approach for clustering, visualization or analysis of *large* data sets [2,5].

The algorithm directing the adaptive process of the SOM mimics the fundamental properties of Hebbian learning using three basic steps after initialization: sampling, similarity matching, and updating. These three steps are repeated until formation of the feature map has completed [4]. The algorithm is summarized as follows:

- Initialization: Choose random values for the initial weight vectors $w_j(0)$, $j=1,2, \ldots,l$, where l is the number of neurons in the map.
- Sampling: Draw a sample x from the input space (without replacement) with a uniform probability.
- Similarity Matching: Find the best matching neuron $i(x)$ at time step n by using the minimum distance Euclidean criterion:
$$i(x) = \arg\min_j \|x(n) - w_j\|, \qquad j=1,2, \ldots,l$$
- Updating: Adjust the weight vectors of all neurons within the neighborhood, $h_{j,i(x)}(n)$, by using the update formula:
$$w_j(n+1) = w_j(n) + \_(n)h_{j,i(x)}(n)(x(n)-w_j(n))$$
- Neighborhood Update: On completion of an epoch, test application of neighborhood annealing schedule.
- Continuation: Continue with sampling until no noticeable changes in the feature map are observed, i.e. $\Delta w(t) \_ 0$.

Furthermore, in this work, SOMs are of particular interest not only for their topological ordering property, feature selection and density matching but also, on account of their ability to provide approximations of the input in a lower dimensional space [4]. In other words, the SOM acts as an encoder to represent a large input space by finding a smaller set of prototypes.

This viewpoint forms the theoretical justification of using the SOMs as encoders, which corresponds to the motivation behind this work. Hence, a hierarchical SOM architecture is used to map the (large) input space to a (smaller) set of prototypes independent from the noise properties. Moreover, the hierarchical architecture of this system makes use of the encoding-decoding model by building up an understanding of PDS by first considering the relationships between characters (the first-level SOMs), then words (the second-level SOMs), and finally to word co-occurrences (the third-level SOMs).

III. Architecture Overview

To achieve the above objective, the framework of figure 1 is followed, where the core of the approach is to automate the identification of typical MRDx characteristics, i.e. good approximation of the input space.

Steps to achieve data pre-processing and reduction are driven by the needs of the pattern discovery component. In this case, as mentioned above, pattern discovery employs a three level hierarchical SOM architecture. Therefore, the significance of effective pre-processing before presentation to the SOM cannot be over-emphasized. PDS must be pre-processed in such a way as to facilitate the training of an unsupervised learning mechanism. The way in which this is performed is unique to this work.

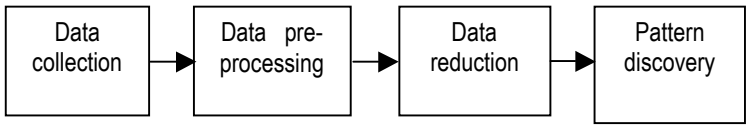

Figure-1: System Overview

A typical pre-processing phase for a document (PDS) uses the Vector Space Model. There are two main parts to the vector space model: Parsing and Indexing [1, 5, 6, 7]. Parsing converts text into a succession of words. The words are filtered using a basic Stop-List of common English words, which do not significantly contribute to the discrimination of a document (such as "the", "it", etc.). A limited stemming algorithm is then applied to the remaining words to reduce plural words to their singular form. In the indexing part, each document is represented in the Vector Space Model where the frequency of occurrence of each word in each document is recorded. These values are then generally weighted using the Term Frequency (TF) multiplied by the Inverse Document Frequency (IDF), following Shannon's information theory as shown in equation (1). Once the set of document vectors has been created, techniques from pattern discovery are employed to form the overall classification.

$$P_{ij} = tf_{ij} \, . \, log \, (N/df_i) \qquad (1)$$

P_{ij} – Weight of term t_j in document d_i
tf_{ij} – Frequency of term t_j in document d_i
N – Total number of documents in collection
df_i – Number of documents containing term t_j

In this work, although the classical parsing method described above is performed, a different indexing scheme is used in order to provide the basis for considering additional features that can hardly be accounted for in usual term-based approaches [8]. For example: the location of a word/term within a document. In this scheme, a PDS is summarized by its words and their frequencies (TF) in descending order, but equation (1) is not used. Once such a list is formed, the words with a TF of three or more in that list are chosen (for each document). Moreover, these words from each document are pre-processed into their characters to train a three level SOM for each MRDx: MRDx specific character encoding, word encoding, and word co-occurrence encoding. By doing so, the authors of this paper aim to minimize the amount of *a priori* knowledge required to overcome the vocabulary differences.

Figure 2 shows the architecture of the system for feature extraction using self-organizing maps. In order to train an SOM to recognize patterns in characters (the first level SOM), the PDS data must be formatted in such a way as to distinguish characters and highlight

the relationships between them. Characters can easily be represented by their ASCII representations. However, for simplicity, they are enumerated by the numbers 1 to 26. The relationships between characters are represented by a character's position, or time index, in a word. For example, in the word *"heart"*: *"h"* appears at time index 1, *"e"* appears at time index 2, *"a"* appears at time index 3, *"r"* appears at time index 4, and *"t"* appears at time index 5. In fact, in the input space the time index is scaled to the range 1 to 26 because Euclidean distance is used as the metric in similarity matching. If one input variable has a greater range than the others it is apt to dominate the map organization because of its impact on measured distances. The process above is repeated for all of the frequently occurring words for each PDS in each MRDx. SOM networks in the same hierarchy only see information specific to one MRDx. It should be noted that it is important to repeat these words as many times as they occur in the PDS. In other words, for a particular PDS, a list of 10 frequently occurring words may occur a combined total of 50 times. Therefore, a list of 50 words is formed to represent the PDS, with the words remaining in the same order as they appear in the PDS.

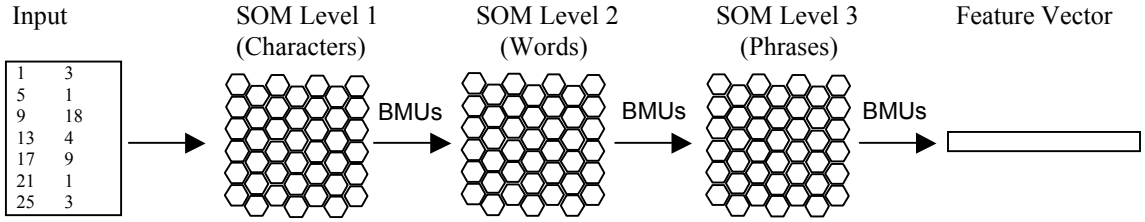

Figure-2: Schematic representation of feature extraction architecture using SOM

When a character and its index are run through a trained first-level SOM, certain neurons are affected more than others, depending on the variations in the statistics of the input distribution, i.e. density matching. Thus, the neurons that are most affected, *Best Matching Units* (BMUs), are used to represent the input space. The BMUs resulting from the first level SOM are used to develop the input matrix to the second level SOM. The second level SOM is trained on these inputs and outputs are again a series of BMUs. This acts as the input data to the third level SOM, trained to recognize phrases or word co-occurrences. Phrases were defined as being a sequence of three words.

The series of words extracted from each PDS are represented in the input matrix to the second-level SOM. Here each row represents a word – where a character with a timing index of 1 identifies the start of a new word – and each column represents a prototype neuron from the first-level SOM. Each of these neurons is a potential BMU, thus if *l* is the number of neurons in the first map, the columns in the input matrix to the second map is also *l*. The three top BMUs are selected for each vector (character) from the first-level SOM and the corresponding column incremented 1 for the first BMU, _ for the second BMU and _ for the third BMU. As an example, if the first character resulted in the three BMUs, 12, 10 and 13, then column 12 would be incremented by 1, column 10 by _ and column 13 by _. This process is repeated for each of the characters in the same word, concatenating the columnar values to form a *word matrix*, which the second level SOM sees as a single pattern.

When a *word matrix* is run through the second-level SOM, input vectors are again reduced to a series of BMUs. Similar to the inputs for the second level, these BMUs are used to develop an input matrix for the level-three SOM. Here the rows of the input matrix represent phrases consisting of three words and the number of columns matches the neurons comprising the second-level SOM. As with the second level, the three BMUs for each word are chosen and values of 1, _ and _ assigned to the proper columns. Each word in the series of words extracted from a PDS marks the start of a new phrase. Thus the number of phrases for a document is calculated by the number of words for that document minus the phrase length plus one.

Finally, the "phrase matrix" is run against the third level SOM to generate an output vector of length *n*, where *n* is the number of neurons in the third level SOM. Again the top three BMUs for each phrase are used to generate values in the columns and these are summed for each phrase found in the document. The resulting vector represents the prototype features for the given document/PDS.

IV. Experimental Study

In this research, the data set required was obtained with the assistance of the Medical Informatics Department of Dalhousie Medical School[2]. This material includes 316 discharge summaries and the corresponding discharge abstracts of patients from Cardiology Services. These PDS, written in English, were from several different manually categorized (by the health records reviewers of the hospital) MRDx. Results detailed below are focused on the feature extraction for the most common MRDx, I25.1, a diagnosis of atherosclerosis, (-seen in more than one third of the cases in the data set). Thus, in total 156 PDSs were used as the training data set (78 for *in-class*, 78 for *out-of-class*), and 48 as the test data set (24 *in-class*, 24 *out-of-class*).

Moreover, the scripts to pre-process the PDS and analyze the results of SOM testing were written in PERL. The scripts to train and test the SOMs were written in Matlab's own scripting language, using the functions of the SOM Toolbox[3]. Visualizations of all data are done using Matlab's built-in facilities and the additional ones provided by the toolbox.

The first-level SOMs are of size seven by five, the second-level of size five by four and the third-level SOMs of size four by four. For the second- and third-level SOMs, input vectors were constructed using the top three BMUs from every character in a particular word, or every word in a particular document, respectively.

As indicated above, two hierarchical SOM systems were implemented; one for the MRDx of I25.1 and another for the MRDx that is a diagnosis other than I25.1. Each system is trained on a subset of the corresponding MRDx data alone. Figure 3 represents the training results for two data collections; the histogram on the left shows documents with a MRDx of I25.1, *in-class*, while the histogram on the right are records with a

[2] The ethics approval was obtained from the QEII Health Sciences Center Research Ethics Board in September 2001.
[3] http://www.cis.hut.fi/projects /somtoolbox.

MRDx that is a diagnosis other than I25.1, *out-of-class*. The x-axis of the graphs are the 16 features extracted automatically and correspond to the number of nodes from the third level SOM; the y-axis shows the frequency with which each map node was identified as the BMU.

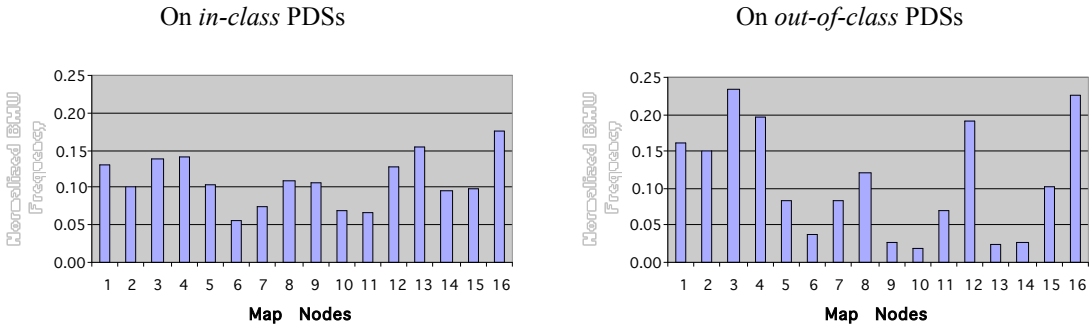

Figure-3: Histograms from the third level SOM of MRDx - I25.1 on the training data set

The histograms show that documents with an MRDx of I25.1 stimulate nodes 16, 13, 4, 3, 1 and 12 most frequently, while those without the target diagnosis stimulate nodes 3, 16, 4, 12, 1, and 2. While there is some overlap, the overall histogram patterns for the PDS sets are distinct where the *in-class* PDS set exhibiting a much flatter shape, while the *out-of-class* PDS show more extreme peaks and valleys. It should be noted that there is substantial difference for the map nodes 3, 4, 9, 12, 13 and 14.

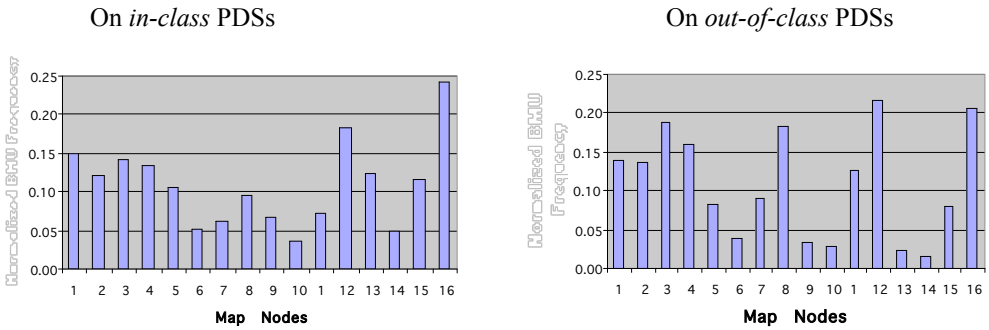

Figure-4: Histogram from the third level SOM of MRDx - I25.1 on the test data set

A test data set was also run on the same SOM series to validate the outputs noted on the training documents and histograms of these results were generated (figure 4). While the histogram of the *in-class* test set shows larger frequencies for nodes 12 and 16, the overall shape is closer to the *in-class* training set than the *out-of-class* training set. The most frequent BMUs for this record set are 16, 12, 1, 3, 4, and 13. In other words, the top six nodes here correspond to the top six nodes for the *in-class* training set. Moreover, node 16 is the single node with the highest frequency as is the case for the *in-class* training set.

On the other hand, the *out-of-class* test set has a BMU sequence of 12, 16, 3, 8, 4, and 1. Overall, the *out-of-class* histogram shows a stronger similarity to the *out-of-class* training

set, with five of the top six neurons matching, although the most excited BMU is 12 in this case, rather than 3. However, due to the random mix of MRDx classifications in the *out-of-class* set, there is not an expectation to see the same level of correlation as that observed with the *in-class* document set.

V. Conclusions

A system has been developed for unsupervised feature extraction that encompasses pre-processing for word features as well as word co-occurrences. By doing so, the authors aim to minimize any *a priori* assumptions regarding suitable word features as well as to discriminate between higher level concepts.

The hierarchical SOM approach to feature extraction extends analysis beyond word frequencies to also consider combinations of words as MRDx indicators. Key medical concepts are often expressed as small phrases – cardiac arrhythmia or myocardial infarction, for example – which may relate directly to assigned diagnosis codes. Histograms generated from the outputs of the SOMs provided a visual indicator as to the potential of this system. There were clear differences in the patterns and sequences of the most frequently excited nodes from the SOMs of the *in-class* and the *out-of-class* systems.

Future work will involve utilization of this system to perform automatic classification using the extracted features. In addition, the authors are also interested in the application of the technique to larger data sets as well as additional target diagnoses from cardiology data. Access to a larger data set of structured discharge summaries should provide sufficient examples to adequately explore new target diagnoses.

Moreover, the authors believe that hospitals need to explore methods to implement well-defined PDSs that use, at least at the departmental level, a standardized set of labels. Ongoing work at Dalhousie is aimed at improving the consistency and construction of discharge summaries. Well-constructed PDSs will expedite automated coding solutions.

Acknowledgements

The authors gratefully acknowledge the financial support of the Nova Scotia Health Research Foundation, Ms. Grace Paterson for her guidance in data collection process, and Dr. Malcolm I. Heywood for his valuable discussions.

References

1. Boone G., Concept Features in Re:Agent, and Intelligent Email Agent, *Proceedings of the 2nd International Conference on Autonomous Agents (Agents '98)*, pages 141–148, New York, 9–13, 1998. ACM Press.
2. Chen H., Schuffels C., Orwig R., Internet Categorization and Search: A Self-Organizing Approach, *Journal of Visual Communication and Image Representation*, Vol. 7, No.1, pp. 88-102, March 1996.

3. Crowley B.L., Zitner D., Farraday-Smith N., Operating in the Dark: The Gathering Crisis in Canada's Public Health Care System, *Atlantic Institute for Media Studies*, http://www.aims.ca/Publications/Dark/Health.pdf. Available 2001.
4. Haykin S., *Neural Networks - A Comprehensive Foundation*, Chapter-9: Self-Organizing Maps, Second Edition, Prentice Hall, 1999, ISBN 0-13-273350-1.
5. Kohonen T., Kaski S., Lagus K., Salojrvi J., Honkela J., Paatero V., Saarela A., Self Organization of a Massive Document Collection, *IEEE Transactions on Neural Networks*, Vol.11, No.3, pp. 574-585, May 2000.
6. Merkl D., Exploration of Document Collections with Self-Organizing Maps: A Novel Approach to Similarity Representation, *Principles of Data Mining and Knowledge Discovery*, pages 101–111, 1997.
7. Merkl D., Lessons Learned in Text Document Classification, *Proceedings of WSOM'97, Workshop on Self-Organizing Maps,* 1997.
8. Sebastiani F., Machine Learning in Automatic Text Categorization, *ACM Computing Surveys*, Vol.34, No.1, pp. 1-47, March 2002.
9. Van Zanten S., Zitner D., Mann K. Proposal entitled, Randomized Controlled Clinical Trial to Improve the Quality of Discharge Summaries of Internal Medicine Residents Using a Discharge Summary Course, 1999.