An Innovative Way for Mining Clinical and Administrative Healthcare Data

Siu Hung Keith Lo and Maiga Chang

School of Computing and Information Systems, Athabasca University, Canada
keithshlo@yahoo.com, maiga.chang@gmail.com

Abstract. A novel method of "predicting" sitter case attribute value is presented in this paper. The method allows users to choose two attributes, seed and target attribute, and to predict the target attribute value of the forthcoming sitter case. The method first retrieves string sequences of the seed attribute according to filters the users set. Then, it finds the words in the sequences and calculates the term frequencies of the words. With the term frequencies, the proposed method uses vector space model to measure the similarity between the testing sequences and the benchmark sequence. At the end, the testing sequence which has highest Cosine similarity value is chosen and the filtering value the method uses to generate the testing sequence is the predicted result. These predicted results allow hospitals to adjust their strategies on resource assignments to better handle patient needs.

Keywords: Healthcare, Regular Expression, Data Mining, Sitter, Hospital Networks.

1 Introduction

Sitter is an external on-call resource hired to "watch" patients who are at risk and need constant supervision. In case something happens to the patient, the sitter informs nurses for intervention. Due to the shortage of in-hospital medical staffs, sitters are hired to free up staffs' time, letting them focus on the jobs which require more skills. It is more cost effective to use sitters to watch patients because

- medical knowledge is often not required, sitters have lower wage than health professionals
- sitters are on-call basis; hospitals call the sitters in and only pay for the particular shifts.

With immense number of patients in public hospitals, reviewing long histories of patient charts and doing analysis can be tedious work. Although statistical reports are being generated to report usages, not much is being done to turn data into real knowledge (i.e., discover patterns and relationships between different clinical and non-clinical elements). Currently, the only feeback on sitter usage are the sitter numbers and dollars spent, some valuable information may still be buried. Such information may be important indications to correlate different clinical and

R. Huang et al. (Eds.): AMT 2012, LNCS 7669, pp. 528–533, 2012.

non-clinical factors, which may be critical to provide both management and practitioners to improve the quality of healthcare.

The purpose of this research is to to analyze sitter usage data in relation to non-nominative patient information and to predict the consequent results directly from previous case sequence without the understanding of the meaning of attributes. Regular expression is applied to the design of the proposed method. The predicted results can be used as a reference to provide healthcare administrators to fine tune staff proportion to better respond patient needs or/and adjust certain procedures to carry out treatments more effectively.

2 Literature Review

In healthcare setting, a lot of information about patient episodes are recorded in various systems. Reports are being generated but they mostly contain only counts, sums and groupings of collected data. Although some manipulations are being done to those reports to facilitate data representation, they are mostly visual appeals or pivot tables, which do not necessarily provide more knowledge or discovery of new information.

According to Fayyad, Piatetsky-Shapiro and Smyth [1], the tremendous amount of data collected and stored in large and numerous data repositories has far exceeded human's ability for comprehension without machine aided analysis. It is almost unavoidable to have data entry errors or inconsistencies in huge data sets. With data mining techniques, outliers can be spotted out and further analysis can be done to determine if those are erratic entries.

Many meaningful patterns can be analyzed and extracted from regular expressions [2]. Regular expression is a metalanguage that describes finite-state automata used for string pattern recognition [3]. It is also a way of describing complex patterns in texts. It has been used to extract information in biomedical field and provided an alternative approach to do complex semantic parsing [9][10][11]. Its advantage is to use shorter and simpler representation for presenting long sequences which contains repeated patterns.

Interesting patterns can be discovered by the recommender system and may assist healthcare institutions to alert health practitioners about some higher risk patients and to find out the reasons of why some patients require higher cost of care and length of stay. For example, some patients with aggressive behaviors can be related to side effects of certain treatments and medications.

3 Clinical and Administrative Healthcare Data

Clinical administrative data such as sitter data is not being used sufficiently. Without discorvering knowledge with data mining techniques, a lot of information may still be hidden. With only numbers showing on statistical reports, more complex questions about patient care cannot be answered. For example, there may be relationships between gender, culture, length of stay, diagnosis, and locations that can affect patient's need for sitters.

According to the hospital's guidelines, no sitter orders can last for more than one entire shift. For patients who need sitter supervision for more than one shift, additional orders must be placed. In every order, the hospital must choose a primary reason from a pre-defined list to explain why the patient requires sitter's further supervision. The reasons can be one of the followings: Agitation, avoid use of 4-point restraints, avoid use of other type of restraint, avoid use of posey vest/jacket, away without leave, behavior problem, constant observation in 4-point restraints, Delirium, Dementia, Disorientation, eating disorder, Psychosis, risk of falls, Suicidal, Trauma, violent, youth protection, and other. The data analyzed by this research includes sitter order's date, department's mission (a.k.a. sub-division), sitter's shift (Day, Evening, Night), primary reason (patient's problem), units that placed the order, health professional who placed the order, health professional (supervisor level) who authorized the order, patient's medical record number, patient's family and first name, patient's gender, patient's primary spoken language, and patient's bed number and location.

The sitter system depends on the hospital's admission, discharge and transfer (ADT) system to get more detailed patient information. Other than the medical record card number and basic information about the patients, the sitter system does not store any other patient specific information but the sitter cases. With data consolidation between the sitter system and hospital's ADT system, it is able to provide the research team the following anonymous data like patient's date of birth (and it allows us to calculate the patient's age), gender, marital status, preferred language, municipality, diagnosis, admission type, admission and discharge date (and it allows us to calculate the length that the patient stays at hospital), and discharge location.

4 Regular Expression Based Data Mining

The research team uses regular expression to summarize and present sitter cases' attribute sequence for a time period. The proposed method can then predict the attribute value that the forthcoming case may have by expanding the regular expression presented the particular cluster, i.e., the regular expression can be seen as a sort of deduction rules. Furthermore, the method applies vector space model to calculate the similarity of two regular expressions. The closer two regular expressions are similar to each other may imply that two sequences may have hidden relationships. Therefore, the method is capable of using an attribute's regular expression to predict the follow-up value of the other attribute.

The objective of the proposed method is to predict the attribute value of forthcoming sitter case based on the attribute values of past sitter cases. Due to all sitter cases have its date and shift stamp, they can be seen as sequential records. Sitter cases after data pre-processing consist of multiple attributes, as Table 1 lists.

Table 1. Sitter cases

Mission	Site	Shift	Reason	Age	Gender	Marital Status	Lang	Adm Type	Length of stay	Discharge Location
Surgery	RVH	Night	Away without Leave	70-79	M	SINGLE_ADULT	French	ER	20-29	Home
ER	RVH	Day	Disorientation	60-69	F	SINGLE_ADULT	French	Stretcher	0-9	Hospital
ER	RVH	Evening	Agitation	70-79	F	SINGLE_ADULT	French	Stretcher	0-9	Hospital
ER	RVH	Night	Disorientation	50-59	F	SINGLE_ADULT	French	Stretcher	0-9	Hospital
Medicine	MGH	Night	Suicidal	80-89	M	MARRIED_ADULT	English	ER	0-9	Home

The proposed method uses regular expression to summarize the values of particular attribute in the case sequence. Before the regular expression can be applied, we need to find a string sequence to represent the cases. The particular attribute is called "seed" which is the attribute the user wants the method to use the seed to predict another attribute's forthcoming value. To facilitate the representation of the sequence elements, a single alphabet index is being used to represent each attribute value. For example, take Reason attribute in Table 1 as the seed, the sequence, EJAJO, is being produced if the following codes are assigned to represent different reasons: E for Away without Leave, J for Disorientation, A for Agitation, and O for Suicidal (O).

From searching a particular attribute and value in the dataset, a string sequence of the chosen seed can be found. A sequence may be similar to some other sequences found by searching for different attributes and values in the dataset. The similarity of two sequences may imply that the follow-up attribute values in two sequences (i.e., the future attribute values predicted from the expansion of the sequence) may have hidden relationship. The proposed method uses the sequence similarity to discover relationships between the values of seed and target attribute. In other words, we assume that the symbolic sequence of an attribute's values may contain hints to reveal another attribute's values. For example, a series of sitter reasons (i.e., the seed attribute) can be used as a predictor to predict length of stays (i.e., the target attribute).

The method does the following steps to predict target attribute's value:

1. deciding seed attribute–seed attribute can be chosen by users;
2. deciding target attribute–target attribute can be chosen by users;
3. filtering the dataset with particular conditions (i.e., specific attributes and values)–the conditions can be chosen by users;
4. generating string sequence (i.e., the benchmark sequence) of the seed attribute values from the filtered dataset;
5. generating string sequences (i.e., the testing sequences) of the seed attribute values from the filtered dataset by using all possible values that the target attribute has as additional filtering criteria;
6. finding words of different lengths for all sequences include the benchmark and the testing sequences;
7. calculating the similarity values for each testing sequences from the benchmark sequence;
8. and, choosing the testing sequence which has highest similarity and use its value as the predicted value for the target attribute.

The proposed method uses word matching technique to determine whether two sequences are similar. A word is a repeated character sequence. A word finding engine within regular expression approach has been developed to find out possible words in different lengths. Once the words are found, they are stored in a dictionary.

By calculating the term frequencies of each sequence and convert them into vector space, with normalized vectors of term frequencies, it can be applied to see how close a string testing sequence is to the benchmark string sequence. Cosine similarity measure [4] has been widely used in clinical analysis to compare sequences generated by data collection tools with timestamps [5][6][7]. It has also been proven to be a robust metric for scoring the similarity between two strings, and it is increasingly being used in text mining related queries [8].

5 Evaluation and Discussion

The research team evaluates the proposed regular expression based data mining method with the data which consists of all the sitter usages within the hospital network, consisting of five hospitals (4 adult sites and 1 child & adolescent site), for the entire years of 2008, 2009 and 2010. To evaluate the accuracy of the prediction results suggested by the proposed method, we compare the predicted results with the known records existed in the dataset. In general, results are quite promising with fair accuracies.

The proposed innovative method help users predict the target attribute value of a forthcoming sitter case based on their chosen seed attributes and criteria. The method doesn't need to know the meanings of attributes and to do complicate calculations like information value and entropy. It simply generates string sequences, finds the words in the sequences, and measures the Cosine similarity a testing sequence has with the benchmark sequence. At the end, the method gives the user its prediction value for the target attribute with the filtering value used to generate the most similar testing sequence against the benchmark sequence.

Such prediction method is important to hospitals. The administrative personnel can prepare the hospital ready for the potential sitter requests, moreover, they can allocate necessary resources like beds and medical professionals with particular skills for the forthcoming patients. The proposed method can also be used to do prediction for the dataset from other disciplines and areas, as long as the dataset is sequential and the attributes used for seed and target attributes are categorical or can be transformed to categorical attributes.

Acknowledgement. The authors wish to thank the support of Athabasca University and the Academic Research Funding.

References

1. Fayyad, U., Piatetsky-Shapiro, G., Smyth, P.: The KDD Process for Extracting Useful Knowledge from Volumes of Data. Journal of ACM Communications 39(11), 27–34 (1996)

2. Lin, C.-H., Hsiao, H.-S.: Hierarchical State Machine Architecture for Regular Expression Pattern Matching. In: 19th ACM Great Lakes Symposium on VLSI, Boston, MA, USA, pp. 133–136 (2009)
3. Jurafsky, D., Martin, J.-H.: Speech and Language Processing: An Introduction to Natural Language Processing. In: Computational Linguistics and Speech Recognition. Prentice Hall, Upper Saddle River (2000)
4. Zhao, Y., Karypis, G.: Evaluation of hierarchical clustering algorithms for document datasets. In: 11th International Conference on Information and Knowledge Management, McLean, VA, USA, pp. 515–524 (2002)
5. Augustyniak, P.: Optimal Coding of Vectorcardiographic Sequences Using Spatial Prediction. Journal of IEEE Transactions of Information Technology in Biomedicine 11(3), 305–311 (2007)
6. Bratsas, C., Hatzizisis, I., Bamidis, P., Quaresma, P., Maglaveras, N.: Similarity Estimation among OWL Descriptions of Computational Cardiology Problems in a Knowledge Base. Journal of IEEE Computers in Cardiology 32(5), 243–246 (2005)
7. Chen, C.-M., Hong, C.-M., Huang, C.-M., Lee, T.-H.: Web-based Remote Human Pulse Monitoring System with Intelligent Data Analysis for Home Healthcare. Cybernetics and Intelligent Systems, 636–641 (2008)
8. Subhashini, R., Kumar, V.J.S.: Evaluating the Performance of Similarity Measures Used in Document Clustering and Information Retrieval. In: 1st International Conference on Integrated Intelligent Computing, Bangalore, India, pp. 27–31 (2010)
9. Grishman, R.: Information Extraction: Techniques and Challenges. International Summer School on Information Extraction: A Multidisciplinary Approach to an Emerging Information Technology, Rome, Italy, pp. 10–27 (1997)
10. Mutalik, P.-G., Deshpande, A., Nadkarni, P.-M.: Use of general-purpose negation detection to augment concept indexing of medical documents. Journal of the American Medical Informatics Association 8(6), 598–609 (2001)
11. Chapman, W.-W., Bridewell, W., Hanbury, P., Cooper, G.-F., Buchanan, B.-G.: A simple algorithm for identifying negated findings and diseases in discharge summaries. Journal of Biomedical Informatics 34(5), 301–310 (2001)