Mining Clinical Pathway Based on Clustering and Feature Selection

Haruko Iwata, Shoji Hirano, and Shusaku Tsumoto

Department of Medical Informatics, School of Medicine, Faculty of Medicine
Shimane University
89-1 Enya-cho Izumo 693-8501 Japan
{haruko23,hirano,tsumoto}@med.shimane-u.ac.jp

Abstract. Schedule management of hospitalization is important to maintain or improve the quality of medical care and application of a clinical pathway is one of the important solutions for the management. This research proposed an data-oriented maintenance of existing clinical pathways by using data on histories of nursing orders. If there is no clinical pathway for a given disease, the method will induce a new clinical care plan from the data. The method was evaluated on 10 diseases. The results show that the reuse of stored data will give a powerful tool for management of nursing schedule and lead to improvement of hospital services.

Keywords: temporal data mining, clustering, clinical pathway, hospital information system, visualization.

1 Introduction

Twenty years have passed since clinical data were stored electronically as a hospital information system (HIS)[7,1,8]. Stored data give all the histories of clinical activities in a hospital, including accounting information, laboratory data and electronic patient records. Due to the traceability of all the information, a hospital cannot function without the information system. All the clinical inputs are shared through the network service in which medical staff can retrieve their information from their terminals [4,8].

Since all the clinical data are distributed stored and connected as a large-scale network, HIS can be viewed as a cyberspace in a hospital: all the results of clinical actions are stored as "histories". It is expected that similar techniques in data mining, web mining or network analysis can be applied to the data. Dealing with cyberspace in a hospital will give a new challenging problem in hospital management in which spatiotemporal data mining, social network analysis and other new data mining methods may play central roles[6,1]. [1] This paper proposes a data mining method to maintain a clinical pathway used for schedule management of clinical care. Since the log data of clinical actions and plans are

[1] Application of ordinary statistical methods are shown in [10,11].

K. Imamura et al. (Eds.): BHI 2013, LNAI 8211, pp. 237–245, 2013.
© Springer International Publishing Switzerland 2013

stored in hospital information system, these histories give temporal and procedural information about treatment for each patient. The method consists of the following four steps: first, histories of nursing orders are extracted from hospital information system. Second, orders are classified into several groups by using a clustering method. Third, by using the information on groups, feature selection is applied to the data and important features for classification are extracted. Finally, original temporal data are split into several groups and the first step will be repeated. The method was applied to a dataset extracted from a hospital informaiton system. The results show that the reuse of stored data will give a powerful tool for maintenance of clinical pathway, which can be viewed as data-oriented management of nursing schedule.

The paper is organized as follows. Section 2 briefly explains background of this study. Section 3 gives explanations on data preparation and mining process. Section 4 shows empirical evaluation of this system on the data extracted from a hospital information system. Section 5 discusses the method and its future perspective. Finally, Section 6 concludes this paper.

2 Background

2.1 Clinical Pathway

Since several clinical actions should be repeated appropriately in the treatment of a disease, schedule management is very important for efficient clinical process[5,12]. Such a style of schedule management is called a clinical pathway. Such each pathway is deductively constructed by doctors or nurses, according to their experiences. For example, Table 1 illustrates a clinical pathway on cataracta in our university hospital. The whole process of admission will classified into three period: preoperation, operation and post-operation. The preoperation date is denoted by -1 day, and operation date is by 0 day. BT/PR denotes Body Temporature and Pulse Rate, BP denotes Blood Pressure.

Table 1. An Example of Clinincal Pathway

Preoperation	Operation	Postoperation				
-1day	0day	1day	2day	3day	4day	5day
BT/PR BP	BT/PR BP	BT/PR BP	BT/PR BP	BT/PR BP	BT/PR BP	BT/PR BP
	Nausea Vomitting	Nausea Vomitting	Nausea Vomitting	Nausea Vomitting	Nausea Vomitting	Nausea Vomitting
		Coaching	Coaching	Coaching	Coaching	Coaching
	Pain	Pain	Pain	Pain	Pain	Pain
Preoperation Instruction						

Notations. BT/PR: Body Temprature/Pulse Rate BP: Blood Pressure

2.2 Related Work

There exists no other research which extracts clinical pathway from hospital information system. This study is an extension of our study on data mining in hospital information system [9].

3 Data Preparation and Analysis

3.1 DWH

Since data in hospital information systems are stored as histories of clinical actions, the raw data should be compiled to those accessible to data mining methods. Although this is usually called "data-warehousing", medical data-warehousing is different from conventional ones in the following three points. First, since hospital information system consists of distributed and heterogeneous data sources. Second, temporal management is important for medical services, so summarization of data should include temporal information. Third, compilation with several levels of granularity is required. Here, data-warehousing has three stages: For hospital service, we compile the data from heterogeneous datasets with a given focus as the hospital information system (HIS) . Then, from HIS, we split the primary data warehouse (DWH) into two DWHs: contents DWH and histories DWH. Then, by using an algorithm shown in Algorithm 1, a temporal dataset for the number of orders will be made as secondary DWH. Data mining process is applied to the generated data sets from this DWH.

Algorithm 1. Data Preparation

Input: L_p= List of Patients for a given Disease
Output: List of *Counter*
while $L_p \neq \varnothing$ **do**
  $Pt \leftarrow car(L_p)$
  Pick up the data for Pt
  $D_a \leftarrow$ data of admission
  $D_d \leftarrow$ data of discharge
  for $i = 0$ to $D_d - D_a + 1$ **do**
    $List \leftarrow$ List of Nursing Orders for $D_a + i$
    while $List \neq \varnothing$ **do**
      $Order \leftarrow car(List)$
      $Counter[i, Order] = +1$
      $List \leftarrow cdr(List)$
    end while
  end for
  $L_p \leftarrow cdr(L_p)$
end while
Return List of *Counter*

3.2 Mining Process

Except for the basic process, we will propose temporal data mining process, which consists of the following three steps, shown in Algorithm 2 We count temporal change of #orders per hour or per days in the second DWH. Then, since each order can be viewed as a temporal sequence, we compare these sequences by calculating similarities. Using similarities, clustering[3], multidimensional scalingMDS, and other methods based on similarities are applied. In this paper, all the analysis is conducted by R2-15-1.

Algorithm 2. Mining Process

```
procedure MINING_PROCESS(Level_v, Level_h, List of Orders
Tables of Number of Orders(Level_v,Level_h)
    L_o ← List of Orders
    T_o ← Tables of Number of Orders(Level_v,Level_h)
      Sim_mat(Level_v, Level_h)
        ← Calculate_similarity_matrix(L_o)
      Labels(Level_v, Level_h)
        ← Clustering(Sim_mat(Level_v, Level_h))
      Apply feature selection methods to T_o
        with Labels(Level_v, Level, h)
                                          ▷ Feature: each date
    Split T_o with the values of Features into T_o[1] ··· T_o[n]
    if n > 1 then
       for i = 1 to n do
          Newlevel_v ← Level + 1
          Table(Newlevel_v, i) ← T_o[i]
          Mining_Process(NewLevel_v, i, L_o,
            Table(Newlevel_v, i))
       end for
    end if
    Return Labels(Level_v, Level_h)
end procedure
```

3.3 Clinincal Pathway Maintenance Process

Algorithm 3 shows the process for maintenance of a clinical pathway. For all the elements in the outputs of the mining process, each order is evaluted by some given function, and if the evaluted value is larger than a given threshold, this order is included in a list of orders during a given period. An evaluation function is provided before the process, the most of a simple one is an averaged frequency of the order during the period. In this study, we use this evaluation function for analysis.

Algorithm 3. Construction of Clinical Pathway

```
procedure CONSTRUCTION_PROCESS(Level_v, Level_h)
    List ← Labels(Level_v, Level_h)
    while List ≠ ∅ do
        Order ← car(List)
        if Evaluation(Order) > Threshold then
            for all attr(Level_h) do
                Append Order into
                List_pathway(Level_v, attr(Level_v, Level_h))
                                        ▷ attr(Level_v, Level_h): List of Dates
            end for
        end if
        List ← cdr(List)
    end while
    Return List_pathway(Level_v, attr(Level_v, Level_h))
end procedure
```

3.4 Similarity

After the construction of clinical pathway, we can calculate similarity between existing pathway and induced one. To measure the similarity, several indices of two-way contigency tables can be applied Table 2 gives a contingency table for

Table 2. Contingency Table for Similarity

		InducedPathway		
		Observed	*Not Observed*	Total
ExistingPathway	*Observed*	a	b	$a+b$
	Not observed	c	d	$c+d$
	Total	$a+c$	$b+d$	$a+b+c+d$

Table 3. Basic Statistics of Datasets extracted from Hospital Information System

	Path Application	Total Cases	#Used Nursing Orders	Min of Stay	Median of Stay	Max Length of Stay
Cataracta (bilateral)	Yes	168	89	4	4	16
Glaucoma	No	156	91	3	3	21
Cataracta (lateral)	Yes	127	78	3	3	6
Lung Cancer (with operation)	Yes	107	170	5	7	50
Brain Infarction	No	106	209	1	12	> 70
Detached Retina	No	88	125	5	14	21
Bladder Cancer	Yes	85	105	4	5	18
Patella & Knee Injury	No	80	67	3	17	44
Lung Cancer (without operation)	No	77	135	1	6	34
Cholangioma	No	68	113	1	6	55

Table 4. Experimental Evaluation on Data extracted from Hospital Information System

	#Intervals	Major Three Intervals			#Used Nursing Orders			Optimal
		Group 1	2	3	1	2	3	Length of Stay
Cataracta (bilateral)	4	0	1 to 3	4	9	5	3	4
Glaucoma	3	-1, 3 to 5	0 to 2	others	7	7	14	3
Cataracta (lateral)	4	-1	0	1,2	3	2	9	3
Lung Cancer (with operation)	8	1	2, 3	4 to 7	7	13	10	7
Brain Infarction	5	2, 10 to 15	3 to 9	17 to 24	34	38	16	15
Detached Retina	3	0 to 12	-2, -1, 13 to 21	others	7	10	19	12
Bladder Cancer	6	0	1	2	11	6	3	2
Patella & Knee Injury	4	0 to 8	9 to 15	-1, 16 to 31	5	10	11	15
Lung Cancer (without operation)	5	2 to 4	3 to 9	1, 10 to 15	14	9	9	7
Cholangioma	4	2 to 10	11 to 15	1, 16 to 31	10	18	10	15

a set of nursing orders used in two pathways. The first cell a (the intersection of the first row and column) shows the number of matched attribute-value pairs. From this table, several kinds of similarity measures can be defined. The best similarity measures in the statistical literature are four measures shown in [3,2].

4 Experimental Evaluation

The proposed method was evaluated on data on nursing orders extracted from hospital information system, which were collected from Apr.1, 2009 to Mar. 31, 2010. The target diseases were selected from 10 frequent diseases whose patients were admitted to the university hospital during this period and where a corresponding given clinical pathway was applied. Table 3 gives the basic statistics of these diseases. For each disease, the proposed method were applied and the outputs were obtained as Table 4. The first, second and third column show the number of data separation, the date used for three major intervals, respectively. And the fourth column shows the estimated optimal length of stay. Three major intervals have the three highest values of information gain, usually, the two intervals neighbor to the interval with complete classification.

If an existing pathway is available, the similarity value between existing and induced pathway was estimated, whose results are shown in Table 5. For a similarity measure, Jaccard coefficient was selected.

The results show that the method is able to construct a clinical pathway for each disease. Furthermore, the best three major intervals suggested the optimal and maximum length of stay, although information on frequency of nursing orders needed to determine the optimal length of stay. For example, in the case of cataracta, the length of stay estimated from three major intervals is 5 days, but with frequency information, the fourth date has smaller frequency of orders, compared with other intervals. Thus, the optimal length will be estimated as 4 days (0,1,2,3).

Furthermore, since a similarity value for bladder cancer is equal to 1.0, the existing pathway captured all the nursing orders needed for clinical care for this disease. On the other hand, similarity value for lung cancer is low, compared with other diseases. Thus, improvement of the pathway can be expected by this method.

In this way, the proposed method can be used for construction and maintenance of clinical pathway, which is equivalent of schedule management of clinical care. This can be viewed as a first step to data-oriented approach into hospital management.

Table 5. Basic Statistics of Datasets extracted from Hospital Information System

	Similarity
Cataracta (bilateral)	0.83
Cataracta (lateral)	0.91
Lung Cancer (with operation)	0.75
Bladder Cancer	1.0

5 Discussion

5.1 Process as Frequency-Based Mining

The proposed method classifies nursing orders used for treatment of patients of a given disease into ones necessary for its nursing care and others specific to the conditions of the patients by using similarities between temporal sequences of the number of the orders. The former can be used to construct a clinical pathway and the latter can be used to risk assessment with detailed examinations of the cases when such nursing orders are applied.

Thus, it can be viewed as an extension of unsupervised frequent pattern mining: frequency plays an important role in classification. By adding temporal nature of nursing orders, frequent orders will be changed during some period: it may be a little complicated when the therapy of a given disease needs tight schedule management. For example, some care should be taken every three days or only the beginning and the end of admission.

For extraction of complicated patterns, the propossed method introduces a recursive table decomposition. After obtaining the labels of clustering results, we use the labels to determine which attributes (dates) are important for clasification. Then, by using the indices of classification power, we split the original table into subtables. Then, we apply again the proposed method to subtables. At most, each subtable only includes one attribute (one date). When all the attributes belong to corresponding subtable, the temporal patterns of nursing orders may be the most complex one. Othewise, according to the granularity of subtables, temporal patterns may have some interesing temporal patterns.

Due to the dependency on frequency, sufficient number of patients is needed for the proposed method to work. If the number of patients is too small, then the method cannot distinguish necessary nursing orders from others. Thus, it will be our future work to extend our method to deal with such a case.

6 Conclusions

In this paper, we propose a general framework on innovation of hospital services based on temporal data mining process. This process can be called similarity-based visualization approach in which similarity-based methods, such as clustering and multidimensional scaling (MDS) and correspondence analysis. We applied the process to datasets of #nursing orders for cases for operation of cataracta where clinical pathway has been introduced. By using Clustering and MDS, we obtained two major groups in the nursing orders: ones were indispensable to the treatment, and the others were specific to the status of patients. Then, in the step for feature selection, the first day of postoperation could be be viewed as a threshold in the original datasets. Thus, periods before and after operation should be dealt as independent datasets. Repeating these steps, we could characterize the temporal aspects of nursing orders, and then found missing information in the existing pathway. This paper is a preliminary approach to data-mining hospital management towards a innovative process for hospital services. More detailed analysis will be reported in the near future.

Acknowledgements. This research is supported by Grant-in-Aid for Scientific Research (B) 24300058 from Japan Society for the Promotion of Science(JSPS).

References

1. Bichindaritz, I.: Memoire: A framework for semantic interoperability of case-based reasoning systems in biology and medicine. Artif. Intell. Med. 36(2), 177–192 (2006)
2. Cox, T., Cox, M.: Multidimensional Scaling, 2nd edn. Chapman & Hall/CRC, Boca Raton (2000)
3. Everitt, B.S., Landau, S., Leese, M., Stahl, D.: Cluster Analysis, 5th edn. Wiley (2011)
4. Hanada, E., Tsumoto, S., Kobayashi, S.: A "Ubiquitous environment" through wireless voice/Data communication and a fully computerized hospital information system in a university hospital. In: Takeda, H. (ed.) E-Health 2010. IFIP AICT, vol. 335, pp. 160–168. Springer, Heidelberg (2010)
5. Hyde, E., Murphy, B.: Computerized clinical pathways (care plans): piloting a strategy to enhance quality patient care. Clin. Nurse Spec. 26(4), 277–282 (2012)
6. Iakovidis, D., Smailis, C.: A semantic model for multimodal data mining in healthcare information systems. Stud. Health Technol. Inform. 180, 574–578 (2012)
7. Shortliffe, E., Cimino, J. (eds.): Biomedical Informatics: Computer Applications in Health Care and Biomedicine, 3rd edn. Springer (2006)
8. Tsumoto, S., Hirano, S.: Risk mining in medicine: Application of data mining to medical risk management. Fundam. Inform. 98(1), 107–121 (2010)
9. Tsumoto, S., Hirano, S., Iwata, H., Tsumoto, Y.: Characterizing hospital services using temporal data mining. In: SRII Global Conference, pp. 219–230. IEEE Computer Society (2012)
10. Tsumoto, Y., Tsumoto, S.: Exploratory univariate analysis on the characterization of a university hospital: A preliminary step to data-mining-based hospital management using an exploratory univariate analysis of a university hospital. The Review of Socionetwork Strategies 4(2), 47–63 (2010)
11. Tsumoto, Y., Tsumoto, S.: Correlation and regression analysis for characterization of university hospital (submitted). The Review of Socionetwork Strategies 5(2), 43–55 (2011)
12. Ward, M., Vartak, S., Schwichtenberg, T., Wakefield, D.: Nurses' perceptions of how clinical information system implementation affects workflow and patient care. Comput. Inform. Nurs. 29(9), 502–511 (2011)