Clustering-based Analysis in Hospital Information Systems

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

Yuko Tsumoto
Department of Fundamental Nursing, School of Nursing,
Shimane University
89-1 Enya-cho Izumo, Shimane 693-8501 Japan
{tsumotoy}@med.shimane-u.ac.jp

***Abstract*—Rapid progress in electronization of hospital information gives an opportunity to realize evidence-based hospital management and services. This paper proposes a clustering-based data mining approach to temporal data in hospital information. The process consists of repetitions of clustering for grouping records and decision tree inducion for feature selection. It will terminate if the clustering results are converged. We evaluated this method to data on order history stored in hospital information system. The results show that the reuse of stored data will give a powerful tool for hospital management and lead to improvement of hospital services.**

***Keywords*-Data mining, Hospital management, Service-oriented computing, Clustering, Decision Tree induction**

I. Introduction

Twenty years have passed since clinical data were stored electronically as a hospital information system. 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.

However, reuse of the stored data has not yet been discussed in details, except for laboratory data and accounting information to which OLAP methodologies are applied. Data mining approach just started ten years ago[1], [2].

In this paper, we first propose a scheme for inovation of hospital services based on data mining. Then, based on this scheme, we applied data mining techiques to data extracted from hospital information systems. The results show several interesting results, which suggests that the reuse of stored data will give a powerful tool to improve the quality of hospital services.

The paper is organized as follows. Section II briefly explains how hospital information system works, which is a background on this study. Section III proposes a hybrid method to analyze temporal histories in hospital information system. Section IV shows the results of visualizaiton of hospital activities by using HIS data. Section V discusses the results with additional analysis. Finally, Section VI concludes this paper.

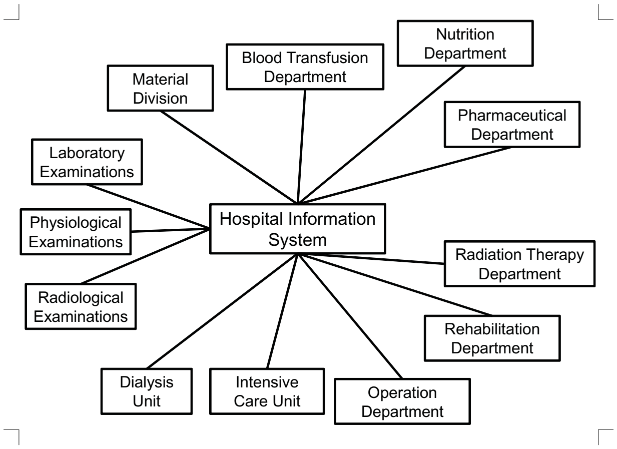
Material Division
Blood Transfusion Department
Nutrition Department
Pharmaceutical Department
Laboratory Examinations
Physiological Examinations
Hospital Information System
Radiation Therapy Department
Radiological Examinations
Rehabilitation Department
Dialysis Unit
Intensive Care Unit
Operation Department

Figure 1. Hospital Information System in Shimane University

II. Background

A. Hospital Information System: Cyberspace in Hospital

On the other hand, clinical information have been stored electronically as a hospital information system (HIS). The database stores all the data related with medical actions, including accounting information, laboratory examinations, and patient records described by medical staffs. Incident or accident reports are not exception: they are also stored in HIS as clinical databases. For example, Figure 1 shows the structure of the HIS in Shimane University Hospital. As shown in the figure, all the clinical inputs are shared through the network service in which medical staff can retrieve their information from their terminals [3], [4].

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

978-1-4577-0371-3/11/$26.00 ©2011 IEEE

and other new data mining methods may play central roles.

B. Basic Unit in HIS: Order

The basic unit in HIS is an "order", which is a kind of document or message which conveys an order from a medical practitioner to others. For example, prescription can be viewed as an order from a doctor to a pharmacist and an prescription order is executed as follows.

1) Outpatient Clinic
2) A prescription given from a doctor to a patient
3) The patient bring it to medical payement department
4) The patient bring it to pharmaceutical department
5) Exceution of order in pharmacist office
6) Delivery of prescribed medication
7) Payment

The second to fourth steps can be viewed as information propagation: thus, if we transmit the prescription through the network, all the departments involved in this order can easily share the ordered information and execute the order immediately. This also means that all the results of the prescription process are stored in HIS.

Figure 2 depicts the workflow of prescription between doctors, pharmacologist, patients and pay desk. For comparison, Figure 3 shows that of injection.

These sharing and storing process, including histories of orders and their results, are automatically collected as a database: HIS can also be viewed as a cyberspace of medical orders.

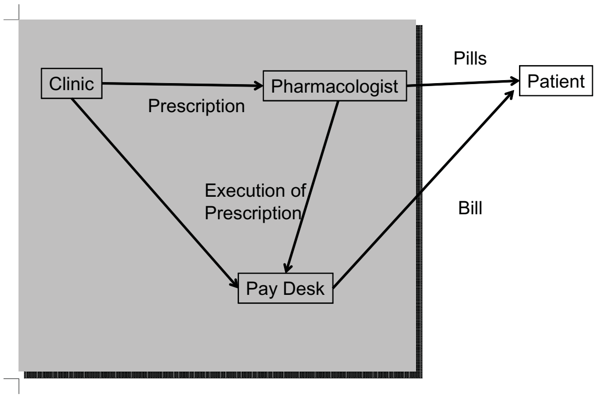

Figure 2. Workflow of Prescription

III. Data Preparation and Analysis

A. Datawarehousing

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 heterogenous 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.

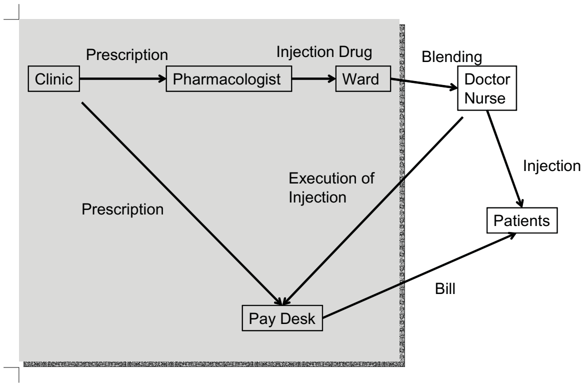

Figure 3. Workflow of Injection

In this paper, we focus on the number of orders to capture temporal global characteristics of clinical activities, whose scheme is given as Fig 4. Here, data-warehousing has two stages: first, we compile the data from heterogenous datasets with a given focus as the first DWH. Then, from this DWH, we split the primary DWH into two secondary DWHs: contents DWH and histories DWH. In this analysis, we focus on the latter DWH and we count the number of orders with a given temporal section, compiled into this DWH. Data mining process is applied to the generated data sets from this DWH.

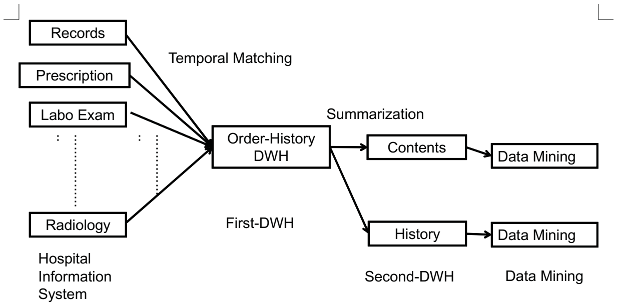

Figure 4. Datawarehousing

B. Analytic Process

Figure 5 depicts a proposed mining process which consists of the following three steps. first, a clustering method is applied to the whole data with all the attributes Then each record is labeled according to the grouping obtained. Secondly, decision tree induction is applied to the labeled data where the new labels are regarded as a target concept. This process can be viewed as a feature selection, and we select features from the derived tree. Then, thirdly, clustering is applied to the data with selected features. If the clusters obtained shows the same classification as the previous ones, then this process will quit. Otherwise, we repeat from the second step.

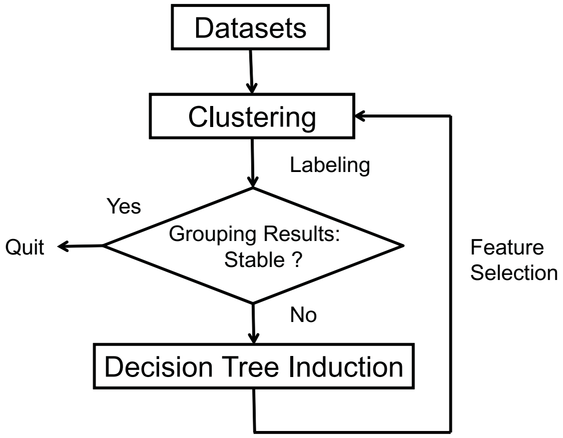

Figure 5. Scheme of Information Reuse

In the analysis below, we selected Ward's method for clustering and C4.5 for decision tree induction.

IV. Results

A. Basic Statistics

Let us show the primitive mining results of HIS. Table I shows the averaged number of each order during the same period. Although these values do not remove the effects of holidays, all the characteristics reflect those shown in Figure 1. Patient records and Nursing cares are the major part of orders (39%) and prescription, reservation of clinics, injection are top three orders in the hospital.

Table II gives the statistics of time differences between ordered and performed date. For example, laboratory examinations are performed 28.55 days (averaged) and 14 days (median) after they are ordered.

Although the above tables overview the total behavior of the hospital, we can also check the temporal trend of each order as shown in Figure 6 and 7. The former figure depicts the chronological overview of the number of each order from June 1 to 7, 2008, and the latter shows that of June 2, 2008. Vertical axes denote the aveeraged number of each order, classfied by the type of orders. Horizonal axis give each time zone. The plots show the characteristics of each order.

Table II
Time Differences between Ordered and Performed Dat (Oct, 2010)

	Average (Days)	Median (Days)
Prescription	0	0
Laboratory Exam.	28.55	14
Physiological Exam.	29.97	7
Radiology	29.22	9
Transfusion	8.897	6
Pathology	-0.003	0
Injection	9.799	2
Reservation	41.85	28
Nursery	1.397	0
Process	-0.0003 0	
Rehabilitation	0	0

For example, the number of records of doctors has its peak in 11am, which correponds to the peak of outpatient clinic, whose trend is very similar to reservation of outpatient clinic. The difference between these two orders is shown in 1pm to 5pm, which corresponds to the activities of wards.

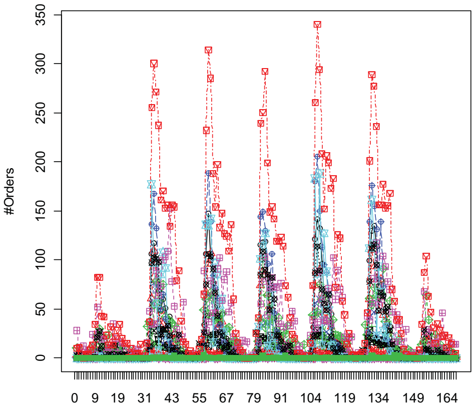

Figure 6. Trends of Number of Orders (June 1 to 6, 2008)

The trends can capture the differences between divisions. Figure 8 and 9 shows those in orders of outpatient clinics of cardiology and rheumatology on Tuesday. Compared with the total trends in outpatient clinic shown in Figure 10, those of cardiology are much closer than those of rhumatology.

These results show that we can measure and visualize the dynamics of clinical activities in the university hospital by exploratory methods. If we can detect some abnormalities different from the usual behavior in these measurements, this may give some knowledge about risks in the clinical activities. Thus, it is highly expected that data mining methods, especially spatiotemporal data mining techiniques play crucial roles in analyzing data in hospital information system and understanding the dynamics of hospital.

B. Clustering with Total Attributes: the first stage

Figure 11 shows the grouping of orders by using Ward's method, in which the metrics are calculated from the chronological trends of the number of total orders in outpatient

Table I
AVERAGED NUMBER OF ORDERS PER DAY (OCT, 2006 TO FEB, 2011)

	Outpatient			Ward		
	Average	Per Patient	Percentage	Average	Per Patient	Percentage
Prescription	403.0095012	0.45733701	11%	259.8551069	0.626198699	7%
Labo Exam	287.9532858	0.326770695	8%	221.3729216	0.533464349	6%
Phys Exam	78.59540776	0.089190425	2%	45.61995249	0.109934937	1%
Radiology	218.7814727	0.248274208	6%	56.35154394	0.135795921	2%
Operation			0%	12.85273159	0.030972506	0%
Transfusion	1.167854315	0.001325286	0%	11.08709422	0.026717674	0%
Meal			0%	186.2929533	0.44892866	5%
Pathology	21.90894695	0.024862372	1%	16.73713381	0.040333136	0%
Injection	84.41409343	0.095793496	2%	469.3942993	1.131146133	13%
Reservations	663.192399	0.752593743	18%	77.46476643	0.186674553	2%
Documents	454.4845606	0.515751141	13%	156.9659541	0.378256473	5%
Nursery	11.08709422	0.012581685	0%	846.3341251	2.039495526	24%
Process	422.7181314	0.479702453	12%	99.04196358	0.238671271	3%
Records	951.9548694	1.080282704	26%	963.0174188	2.320678865	28%
Rehabilitation	3.047505938	0.003458324	0%	3.064924782	0.007385854	0%
In/Out			0%	55.40300871	0.133510141	2%
Total	3602.315123	4.087923541	100%	3480.855899	8.388164698	100%
#Patients	881.2090261			414.9722882		

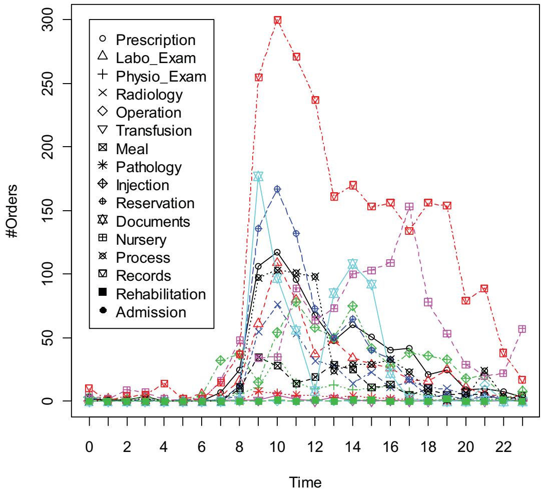

Figure 7. Trends of Number of Orders (June 2, 2008)

clinic during October, 2010. As shown in this figure, the divisions can be classified into two major categories.

C. Decision Tree for the first stage

With the labels obtained, decision tree induction was applied to the dataset. Figure 12 shows the result where only selected attribute is "T11", Tuesday 11pm. Thus, this time zone is the best attribute for classification of two major groups.

D. Clustering with Total Attributes: the second stage

Figure 13 shows the grouping of orders by using Ward's method, in which the metrics are calculated from the selected attribute "T12". As shown in this figure, the divisions can be classified into three major categories.

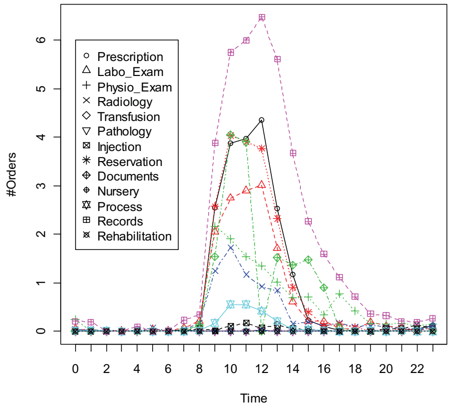

Figure 8. Trends of Number of Orders of Cardiology(2010)

E. Decision Tree for the second stage

With the labels obtained by the second clustering, decision tree induction was applied to the dataset. Figure 14 shows the result where only selected attribute is "T9" (Tuesday 9am) and "T12" (Tuesday 12pm). Thus, these time zones are the best attribute for classification of three major groups.

Since the clustering result obtained by these two attributes is the same as that of the first stage, the process was terminated.

V. DISCUSSION

The clustering method gave three major groups of divisions in the university hospital. The first one is the divisions of surgeon specialities whose patients are concentrated on tuesday. The second one includes other surgeon specialities. The final one is dominated by internist specialities. Thus,

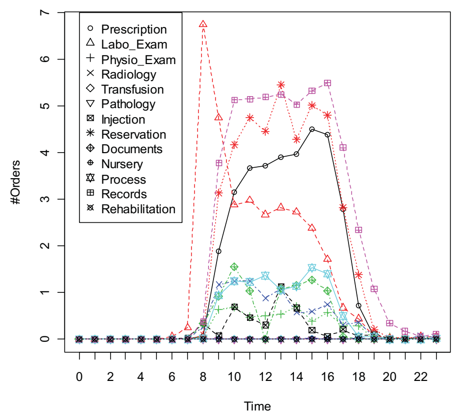

Figure 9. Trends of Number of Orders of Rheumatology(2010)

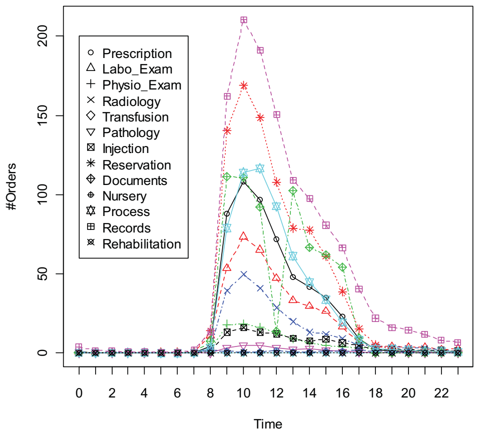

Figure 10. Trends of Number of Orders of Outpatient Clinic(2010)

these results show that the temporal behavior of this university hospital is charaterized by temporal trends of Tuesday morning.

MDS gave additional patterns. Figure 15 shows the two-dimensional mapping obtained by the similarity matrix calculated in the second step, where clearly four groups can be observed. This means that multidimensional scaling gave another axis for grouping: neurological and non- neurlogical specialities. The center of these groups is the third group in the clustering results, around this center, the second group includes the divisions of abdominal diseases and the third group includes the dvisions of neurological diseases. The fourth one is the subset of the first group in the clustering.

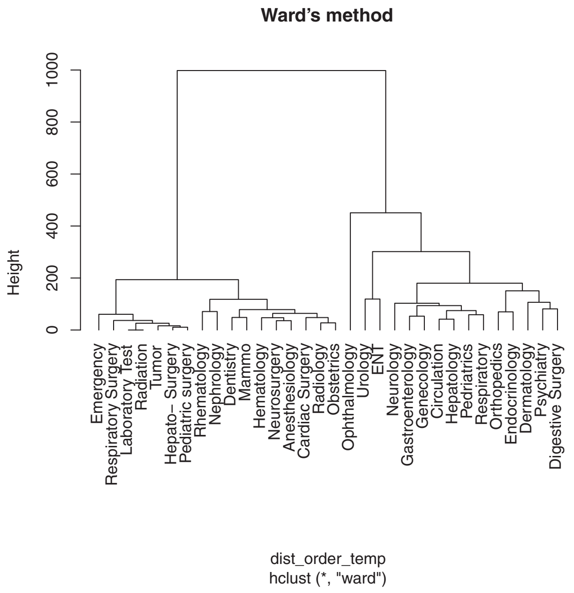

Figure 11. Clustering of Divisions with All the Attributes

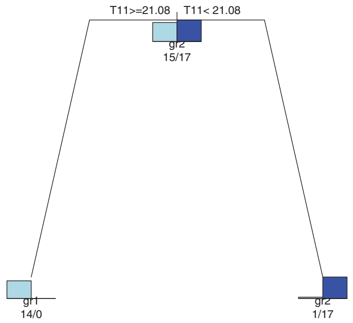

Figure 12. Decision Tree induced by Chronological Patterns of # Total Orders

VI. Conclusions

In this paper, we propose a cluster-based analysis of temporal order histories in hospital information system. The process consists of the following four steps: (1) clustering method is applied to preprocessed data on #orders. (2) By using the labels obtained, features best for classification of the lables are selected by decision tree. (3) clustering is again applied with respect to selected attributes. (4) The above three processes will be repeated until the convergence.

The method was evaluated on data on # total orders calculated from (secondary) DWH. In the first step, Tuesday 11am was selected as the best feature, and the final clustering results showed that all the divisions were classified into three major classes. Then, in the next step, Tuesday 9am and 12pm were selected as the best features.

Three major classes consist of (1) the divisions of surgeon specialities whose patients are concentrated on tuesday, (2) other surgeon specialities and (3) internist specialities. The

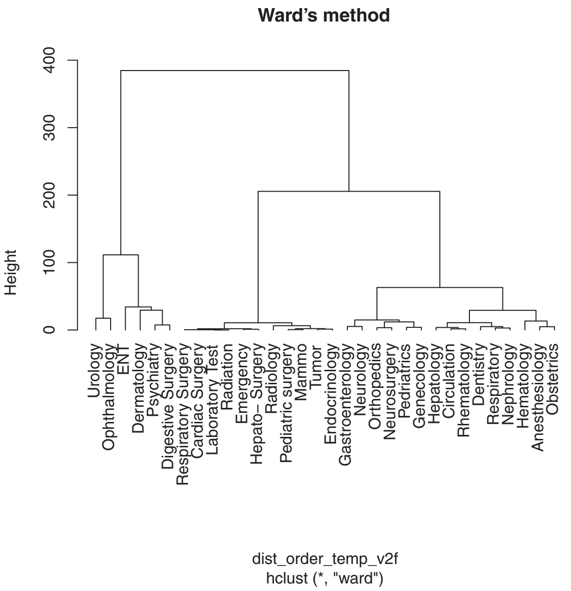

Figure 13. Clustering of Division with T12

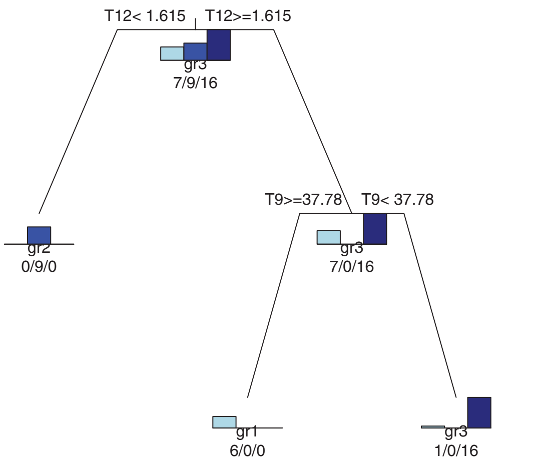

Figure 14. Decision Tree induced by Chronological Patterns of # Total Orders

results obtained by multidimensional scaling also shows another axis for grouping: neurological and non- neurlogical specialities. As a first step, these results show that the reuse of stored data will give a powerful tool for hospital management and lead to improvement of hospital services. This paper is a preliminary approach to data-mining hospital management towards a innovative process for hospital services. More detalied analysis will be reported in the near future.

ACKNOWLEDGEMENTS

This research is supported by Grant-in-Aid for Scientific Research (B) 21300052 from Japan Society for the Promotion of Science(JSPS).

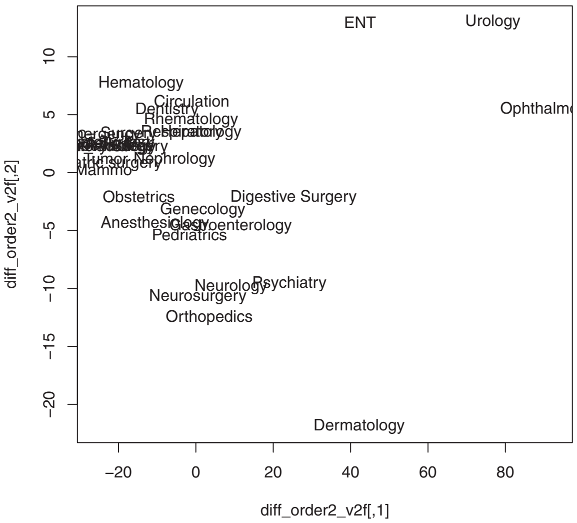

Figure 15. MDS Results of the Second Step

REFERENCES

[1] S. Tsumoto, “Knowledge discovery in clinical databases and evaluation of discovered knowledge in outpatient clinic,” *Information Sciences*, no. 124, pp. 125–137, 2000.

[2] ——, “G5: Data mining in medicine,” in *Handbook of Data Mining and Knowledge Discovery*, W. Kloesgen and J. Zytkow, Eds. Oxford: Oxford University Press, 2001, pp. 798–807.

[3] E. Hanada, S. Tsumoto, and S. Kobayashi, “A hubiquitous environmenth through wireless voice/data communication and a fully computerized hospital information system in a university hospital,” in *E-Health*, ser. IFIP Advances in Information and Communication Technology, H. Takeda, Ed. Springer Boston, 2010, vol. 335, pp. 160–168.

[4] S. Tsumoto and S. Hirano, “Risk mining in medicine: Application of data mining to medical risk management,” *Fundam. Inform.*, vol. 98, no. 1, pp. 107–121, 2010.