Nursing Informatics
U. Gerdin et al. (Eds.)
IOS Press 1997

Data Mining: A Strategy for Knowledge Development and Structure In Nursing Practice

L.R. Eriksen[a], J.P. Turley[a], D. Denton[b] and S. Manning[c]
[a] *Department of Nursing Systems and Technology, University of Texas-Houston School of Nursing, Houston, TX 77030,* [b] *Department of Management Engineering, St. Luke's Episcopal Hospital, Houston, TX 77030,* [c] *Nursing Administration, St. Luke's Episcopal Hospital, Houston, TX 77030*

Data mining is an emerging technique used more widely by the business world than the world of nursing and health care. However, this strategy can be helpful for improving the quality of decision making by clinicians and health care administrators. This paper addresses the concepts and techniques of data mining that could be useful for practicing nurses as well as nurse administrators. Data mining can be an important tool for the development of nursing knowledge and knowledge structures. An example of the use of the technique in an inpatient setting is provided and insights from the process are discussed.

Introduction

The explosive development of high powered computers and storage technology over the last 10 years has made possible the collection and storage of extremely large volumes of data. The large amounts of data have become overwhelming, generating the need for new approaches to turn gigabytes and terabytes of data into useful information and knowledge.[1] Information specialists have developed new techniques and tools to accomplish this. These strategies are emerging from the field of data mining and knowledge discovery. Nurses and other health providers have not begun to use these promising approaches in either the areas of clinical or administrative decision making. In addition, the application of data mining techniques to the structure and development of nursing knowledge has yet to be explored. As computer based patient record systems and taxonomies are developed it will be essential to incorporate into computer based patient records and taxonomies data elements necessary for the development of nursing knowledge.

Concepts

Data mining originated from work done on artificial intelligence as well as the development of data query strategies and tools. Data mining is a process that searches for patterns in a database which can provide information leading to knowledge development about products and processes addressed in the database. Technologies for data mining include neural networks, rule induction, nearest-neighbor analysis, decision trees and data visualization to name a few.[2] Data mining techniques are useful for finding patterns in data which reveal associations or sequences of events not previously known to exist. Collecting useful data for mining is a complicated task. The availability of archived data which has been used for business or clinical systems operations requires extensive extracting, filtering, cleaning and aggregating before it becomes useful for mining.[2] In addition to 'clean data', data mining requires a clear model of expected relationships which may be found in the data. It is necessary to know what problems you want to address to insure the appropriateness of the data model which drives the structure of the data warehouse. Electronic health records will generate large amounts of health related data that will reflect the interventions of many health providers and the response to those interventions will be seen in the patient outcomes.

Figure *1* illustrates the data sources and processes leading to a data warehouse which can then be mined for knowledge. Health care agencies and systems are collecting gigabits or terabytes of operational data which if extracted, cleaned and aggregated can become useful databases for drilling down for 'nuggets' of knowledge.

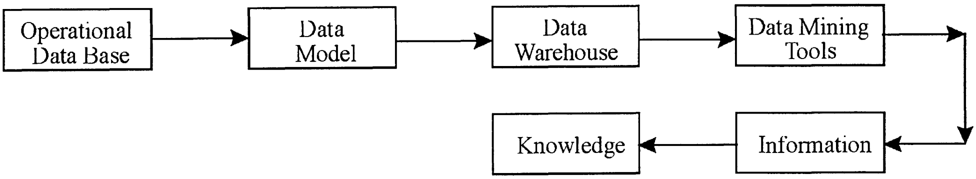

Figure 1. Flow of data from base to knowledge.

Data mining uses different techniques to find useful information in a large database. These techniques include user initiated queries, an automated data mining program or a combination of both. In the user initiated query approach, the user identifies the question to ask of the data base or proposes a relationship model which may be substantiated in the database. An automated data mining program looks for interesting associations or patterns in the database which are not previously identified by the user.[3] These patterns can, in turn, be used to describe past trends and to develop predictive models. The process is iterative in nature[4]. A search result is reviewed by an analyst who interprets the output.

Data visualization is another technique used to display complex data. Some techniques allow the ability to 'fly' through data representations as if you are no longer constrained by gravity and can view the data from many different angles and across time. Viewing from a variety of angles opens new ways of 'seeing' what is not obvious using flat static data representations. Other data visualization techniques use colour and texture to display complex forms which represent the underlying data patterns. Visualization techniques recognize that the human brain is the most efficient organ for the recognition of complex patterns in data. This requires substantial training, as it does in any other field.

Each of the mining techniques is based on different mathematical and statistical manipulations. While it is beyond the scope of this paper to discuss each of the mathematical assumptions of the data mining techniques, users of the techniques must be cognizant of the assumptions and the limitations of each of the techniques before progressing with their use. The statistical technique must be compatible with the model or information type which the analyst is seeking. While the programs which embody these techniques are relatively new on the market, the number of the available programs is expanding exponentially. Indeed many of these data mining programs are being shipped with data warehouse and high end database programs.

Information Advantage summarizes the 15 key capabilities of data mining as: 1) multidimensional view, 2) Pivot/Rotation capability, 3) Intelligent drilling, 4) Cross-dimensional calculation, 5) Dynamic sets, 6) Filters, 7) Decision groupware capabilities, 8) Collapsible browsing, 9) Flexible period definition, 10) Direct access to relational databases,

11) Meta-data, 12) Sparse matrix schema support, 13) Read and write to data warehouse, 14) Query generation and 15) Openness to relational databases.[5] A summery of data mining tools can be found at URL: http://info.gte.com/gtel/sponsored/kdd/siftware.html (November 14, 1996).

The data mining example

In order to experiment with data mining we elected to use a readily available Access™ data base. This data base was developed from a study of nursing activities at a regional medical center hospital. The data base was reviewed and variables of interest selected and saved as our working data base from which iterative explorations were pursued. We also selected KnowledgeSEEKER® from Angoss software as the program to use for our mining.

Data base development: patient classification system project

The original project was undertaken to develop a patient classification system for staffing patient care areas based on patient acuity and patient volume. Two committees were developed. One committee to oversee the coordination of the project and the second to plan and implement the system.

The question of developing an in-house classification system or to purchase a system was the first decision considered. Research into available systems resulted in the decision to develop the system in-house to avoid the specialty of care issues that require the customization of a purchased system. It was also decided that the data collection process would incorporate a new technique of using small, credit card sized barcode readers issued to each staff member to record task or activities performed.

Classification levels were developed including clinical descriptions to distinguish each level of patient acuity. Additionally, a list of activities and tasks performed on a patient care unit was created and definitions developed. Activity lists were further subdivided and color coded into job groupings (i.e. licensed, unlicensed, management, clerical, etc.) to facilitate barcoding and to reduce the number of activities each employee would be required to use. Inservice education programs to train the staff were scheduled one week prior to the beginning of each barcoding study. During these inservices, the staff members were introduced to the activities list, activity definitions, and classification levels with clinical descriptions. A second inservice was held at the start of each barcoding study to familiarize each staff member with the operation of the barcode readers and the use of the scanning booklets. Also covered were the proper procedures for documenting corrections or omissions on the correction sheets.

The staff barcoded bed numbers and activities for each task performed during the course of their shift. An internal clock built into each barcode reader recorded the time and date that each activity was scanned. After each shift, the staff was instructed to place their barcode reader into the battery rechargers / data downloaders. Data obtained from the daily downloads was imported into a spreadsheet for review and correction of duplicate barcodes, errors, missing room numbers and other problems. Reports were printed daily and distributed to each staff member for review and editing.

After the completion of each barcoding study, room numbers were matched to patient names and ID numbers and entered in the database. Classification sheets were then used to associate classification levels to each patient. Completed databases were analyzed and standards developed for each position type by classification level and shift. These standards were used in the development of a computerized classification tool to project the staffing levels based on the patient acuity and patient volume.

KnowledgeSEEKER®

KnowledgeSEEKER® is a data mining tool based on a cluster analysis technique. This cluster analytic technique is unusual in that it combines merging and splitting techniques[6]. This strategy results in multi-way branches or clusters that are grown in the program. A Bonferroni adjustment is made to test the significance levels to accommodate potential type 1 errors. The software program was able to generate both the data tree and IF THEN ELSE rules which could be used in the development of a decision support system.

To run this program we used a Pentium 100 with 32 megs of RAM under Windows NT 3.51. Individual analyses took from several minutes to several hours, indicating the need for a large workstation when using large data sets.

Variables Included in the study were:

PT_CLASS	Patient Classification, a measure of acuity
START_TIME	Time of day an activity was initiated
STOP_TIME	Time of day that an activity was completed
ACTIVITY_T	The amount of time it took to complete an activity
ACTIVITY_D	Calendar day on which the activity was performed
SHIFT_BEGI	Calendar date on which the nursing shift began
JOB_CODE	Job classification code of the person performing an activity
ROOM_NUMBE	Room of patient in which the activity occurred
EMPLOYEE_N	Employee number

While patient name and employee name were collected in the original data, they were purged before any data analysis began.

Data analysis

In preparation for analyzing the data for this study, it was necessary to export the original data in ACCESS 2.0® to an acceptable file structure . Newer versions of KnowledgeSEEKER® directly read ACCESS files. For this first pilot run, the data from a single unit was selected. The data file was in excess of 10 megs. Once the data was in the proper format the program was run in 'automatic' mode. The initial run showed that some of our data was indeed duplicative and was confounding the results. PT_CLASS and CLASS_CODE were both measures of patient acuity. In consultation with those who had designed the original study we dropped CLASS_CODE. The EMPLOYEE_N was also dropped in favor of JOB_CODE as it was the job classification code that we were seeking not the individual practitioner. JOB_CODE was recoded. Originally there were 18 separate job classification codes. The result was too granular. After recoding, JOB_CODE was grouped into licensed personnel and non-licensed The remaining variables were used in further analysis.

Following is a section of a KnowledgeSEEKER® tree that was generated with the Amount of time given to activities as the dependent or predicted value. Each node (or box) on the tree you will see the average time for that node, the standard deviation of the time and the categories that generated the node, e.g. a range of room numbers or a range of stop times. In examining the tree, we found that the average amount of time spent in delivering nursing activities changes by the room in which the activities were delivered and by the start time of the shift on which the activities were performed. While we might assume that the number of activities performed might change by room and by shift, the fact that the amount of time given to an average activity results in some need to further investigate why the combination of room and shift start time seems to impact the amount of time for the delivery of activities. One could assume that nurses would assign patients with a higher acuity or dependency to rooms closer to the nurses' station. But acuity measured as a code did not seem to have any significant relationship in the model. Hence it could be that the acuity or dependency code did not accurately reflect the measure of dependency, or it may well be that there is some other factor or set of factors which impact the amount of time given to nursing activities in the delivery of care. In this case the resulting figures have generated the need for further investigation rather than posing the final solution. However, the model gives us sufficient direction that we know we need to investigate a relationship between certain rooms and the starting time of the nurses who perform the activities.

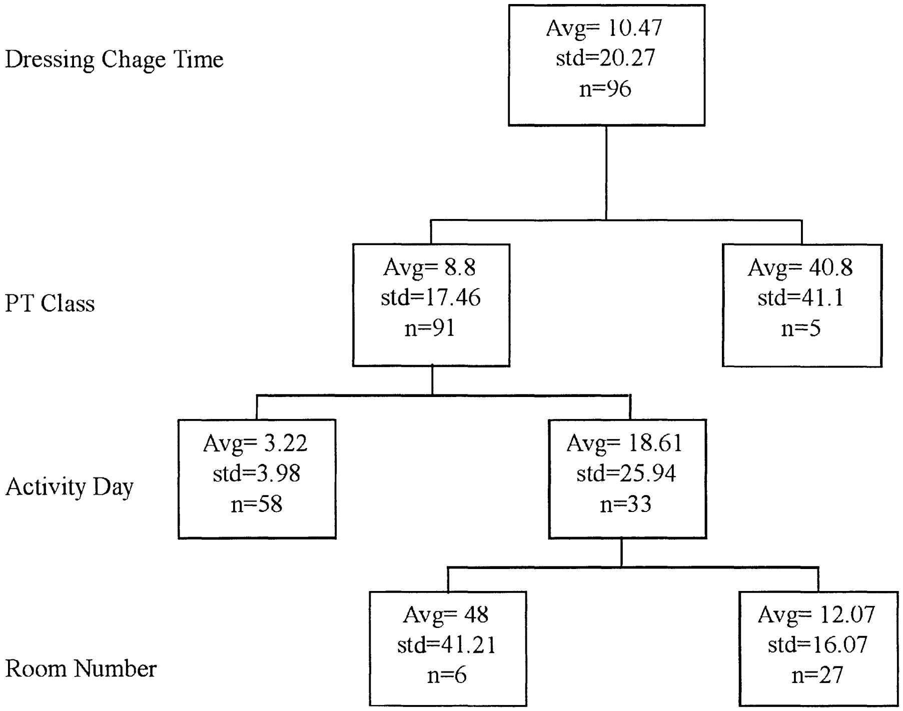

Figure 2. Example of clustering tree.

As we further explore these relationships there may be direction for the structuring of activities, work schedules and level and number of personnel required for providing care on the unit. The clustering approach used by this software proposed a problem and suggested a direction for solution. Unfortunately it is not possible to display the entire tree in any readable form. A partial example of a cluster tree is presented in Figure 2. This analysis led to a number of issues which were addressed in further analyses. This interaction with the automated knowledge discovery process was common in our interaction with KnowledgeSEEKER.®

In another set of analyses, we examined the variables that predicted the amount of time it would take for the performance of each activity. For this analysis, we had to stratify the database so that each subset would include only one nursing activity. The program was then asked to determine the elements affecting the delivery of care for dressing change. In the more detailed analysis of the time it takes to change dressings, a number of issues arise. The first is that one PT_CLASS takes more than 40 minutes on average to change dressings and that within this group the standard deviation is 41 minutes. Even though the number of patients in this group or node is small (n=5), this is a group worthy of some special attention as they are radically different than the other 91 patients considered.

For the remaining 91 we see there are considerable differences in dressing change by the day on which the dressing was changed and the room number of the patient. As in our previous example above, the program has both generated what at first appears to be a difficulty and also gives us some direction in which to seek a solution. What is impressive is that the data mining program seems able to find significant clustering differences even when the size of the nodes are relatively small e.g. nodes of n=5 and n=6 above. In summary, this pilot investigation into the use of data mining software has generated a number of surprises and insights. The use of the data mining software gave a mechanism for understanding a complex data set which had not been amenable to analysis using traditional statistical analyses. The data mining techniques brought to the fore patterns in the analysis which allowed for the classification of groups. In the dressing change example above, the program was able to generate IF..THEN..ELSE rules which could be applied in decision support systems or to further our understanding of the relationships among the variables.

In the overall analysis of the time it takes to perform activities, we were able to see the origin of patterns by room number, time of day and the like. The ability to graphically display the model as a tree/node structure facilitates the understanding of the myriad of complex relationships which are embodied in the data. Data mining's ability to give direction for future analysis and its involvement in an iterative process were far more important than we would have believed before beginning this pilot study. We believe that the type of software we used is an important new tool for nursing and the health professions to better understand the practice of health care and to use a research base to make the deliver of care both more efficient and of higher quality.

The implications of data mining are throughout nursing practice. In the examples given here, a more complex acuity system than currently exists can be developed. Once we know that there are different time allocations an activity, such as a physical exam at different times of the day and/or with different patient groups, then acuity systems can me made to more closely reflect the actual practice of nurses. Once the proper data is included in the taxonomies and databases, the ability to demonstrate outcome effectiveness of different interventions over a period of care becomes a possibility. Data mining is only a technique, admittedly a powerful one, but it will not be effective without the data to drive it and the willingness of nurses to use available data to modify their practice.

References

1. Hedberg SR. The data gold rush. *Byte* 1995;October:83-88.
2. Richman D. Data mining chisels its niche. URL: http:/www.computerworld.com/search/AT-hlm/open/960129SL4mine.html#TOF; January 29,1996.
3. Connor L. Mining for data. URL: http://techweb.cmp.com/cw/021296/tech596.htm#; February 12, 1996.
4. Information Discovery, Inc. URL: http://datamine.inter.net/datamine/dsanal.htm, January,6,1997..
5. Information Advantage. Understanding Multidimensional Analysis: The 15 Keys. Minneapolis, MN: Information Advantage.
6. Angoss Software. KnowledgeSEEKER. URL: http://www.angoss.com/ks/tech.htm; April 9,1996.