Proceedings of the 5th WSEAS International Conference on Applied Computer Science, Hangzhou, China, April 16-18, 2006 (pp776-781)

Intelligent Healthcare Data Analysis using Statistic Data Miner

JIANG B. LIU and YUFEN HUANG
Computer Science and Information Systems Department
Bradley University
Peoria, IL 61625, U.S.A.

Abstract: Nationwide Inpatient Sample (NIS) provides a rich data set of hospital inpatient stays information. In this research we have taken a subset of the NIS data to mine the information of those patients who have the gastro esophageal reflux disease (GERD). The mining process was developed using the Statistica Data Miner (SDM) to establish trends in surgical techniques, patient characteristics and short-term hospital outcomes of GERD in children. These mining results can be used by the health care providers to better service the GERD patients.

Key-words: Statistic Data Mining, Healthcare Data Analysis.

1 Introduction

The Nationwide Inpatient Sample (NIS) [1] is one in a family of databases and software tools developed as part of the Healthcare Cost and Utilization Project (HCUP) [2]. HCUP is a Federal-State-Industry partnership sponsored by the Agency for Healthcare Research and Quality, HCUP data informs decision making at the national, state, and community levels. NIS is a unique and powerful database of hospital inpatient stays. It is the largest all-payer inpatient care database in the United States. It contains data from approximately 7 million hospital stays. NIS can be used to identify, track, and analyze patient data for a better understanding of the health care utilization, access, charges, quality, and outcomes. NIS's large sample size enables analyses of rare conditions, such as congenital anomalies; uncommon treatments, such as organ transplantation; and special patient populations, such as children.

In this research a subset of the NIS data was used to mine the information of those patients who have the gastro esophageal reflux disease (GERD). GERD is common co morbidity for children with chronic conditions and diseases. There are limited non-surgical approaches to control the symptoms of GERD. Pediatric surgical approaches to GERD in recent years have advanced with published reports of excellent results. Advances in surgical techniques with widespread acceptance and utilization results in an increased number of children treated surgically. In spite of technical advances and widespread experience in performing these techniques, co morbid conditions and chronic diseases in children with GERD may still increase the risk of complications and could worsen outcomes. The purpose of our mining is to better understand the trends in surgical techniques, patient characteristics and short-term hospital outcomes of the GERD patients. The healthcare data analysis process used consists of data cleaning, reduction, modeling, and mining, The Statistica Data Miner (SDM) was used to mine the NIS data mart using the clustering analysis techniques.

2 Data Set Reductions, Cleaning, and Modeling

A subset of the NIS data was created to mine the GERD patients. The gigabytes data set was cleaned and reduced to a manageable megabytes data mart using the following process.

a. Unzip the NIS data set and save it as a portable file using SPSS.

b. Clean and Convert the SPSS file to a SDM database using SDM.

c. Reduced the database by creating a data mart with a population of GERD patients who are less than 18 years old and have less than a year hospital stays.

d. Analysis the data mart in SDM.

Figure 1 listed the Data Mart Model. All data in the data base are weighted by NIS in order to compare them nationally. Every time the data mining performs a statistical analysis, association rules, cluster analysis, and chart or graph the program will automatically adjust the data according to the weight given by the NIS.

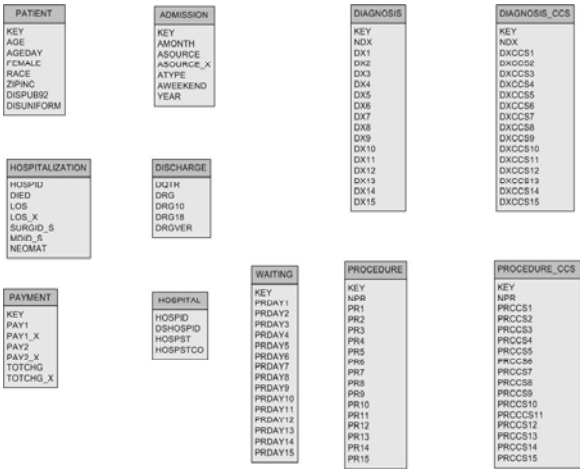

Figure 1: NIS GERD Patient Hospital Stays Data Model.

3 GERD Data Mart Clustering Analysis

The Statistica clustering module is specifically designed to handle large data sets and to enable clustering of continuous and/or categorical variables. It provides the functionality for complete unsupervised learning for pattern recognition. Various cross-validation options are provided that will automatically choose and evaluate a best solution for the clustering problem. This clustering module will be used to mine the GERD data mart.

3.1. K-Means cluster analysis

In this analysis of using k-mean clustering algorithm, this research has selected the primary diagnose (DX1) as the categorical dependent variable, length of hospital stay (LOS) as the continuous predictor variable, and discharge weights that is used to create national estimates for all analyses on the 10% samples (DISCWT10) as the weight variable. The number of clusters and the maximum number of iterations are selected for the optimized analysis. The result is shown in table 1. From the cluster analysis, it can see that the cluster 2 has a higher probability density and higher percentage of occurrence than other clusters.

Centroids for k-means clustering
Number of clusters: 3
Total number of training cases: 544

	LOS	Number of cases	Percentage (%)
1	103.2851	10	1.83824
2	5.5987	445	81.80147
3	27.4264	89	16.36029

Table 1 Summery of 3 clusters for LOS

In next phase, the cluster 2 was focused on to conducting further analysis on the relationship between DX1 and Gender, and DX1 and Race using statistic technique of cross tabulation table and frequency tables.

Relationship between DX1 and Gender

In this cluster, 77.98% of patients having DX1 code, 53081, and female of these patients has slightly higher percentage than male (Table 2).

2-Way Summary Table: Observed Frequencies

	DX1	MALE	FEMALE	Row
Count	53081	198	149	347
Total Percent		44.49%	33.48%	77.98 %
Count	5533	10	1	11
Total Percent		2.25%	0.22%	2.47%
Count	53011	12	8	20
Total Percent		2.70%	1.80%	4.49%
Count	9974	5	4	9
Total Percent		1.12%	0.90%	2.02%
Count	5070	5	4	9
Total Percent		1.12%	0.90%	2.02%
Count	All Grps	260	185	445
Total Percent		58.43%	41.57%	100%

Table 2. 2-way summary table for DX1 and Gender

Therefore, the gender is not the major influence in this cluster but the DX1 code is. The patients with DX1 code 53081 dominated the cluster. The other DX1 codes (5533, 53011, 9974, and 5070) are not significant in this cluster. (Figure 3).

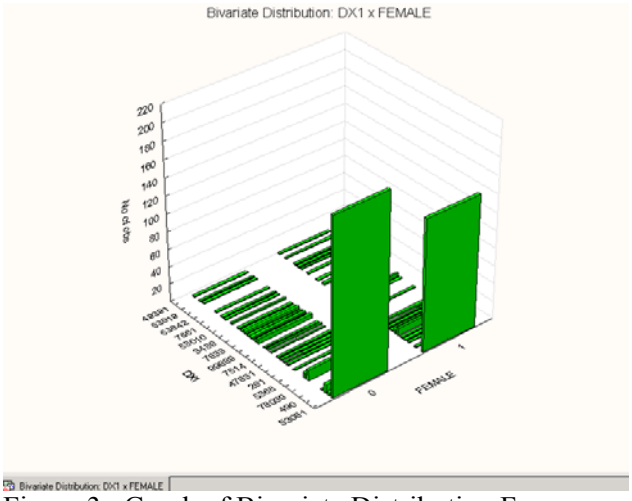

Figure 3. Graph of Bivariate Distribution For DX1 and Gender.

Relationship between DX1 and Race

From Table 2 and Figure 3, it is clearly indicated that the patients with DX1 53081 are the dominated group in the cluster. The reace distribution within the group has been studied. Table 3 listed the detailed observed frequencies for DX1 and Race. In this group, 70.82% of patients are whites, 10.89% are blacks, 13.62% are Hispanics, which are comparable to the overall population in the nation. Further analysis is needed to narrow the patient's data sets to compare with the regional race distribution data. It will be useful to compare the urban patient's distribution with rural patient's distribution to identify the race factor.

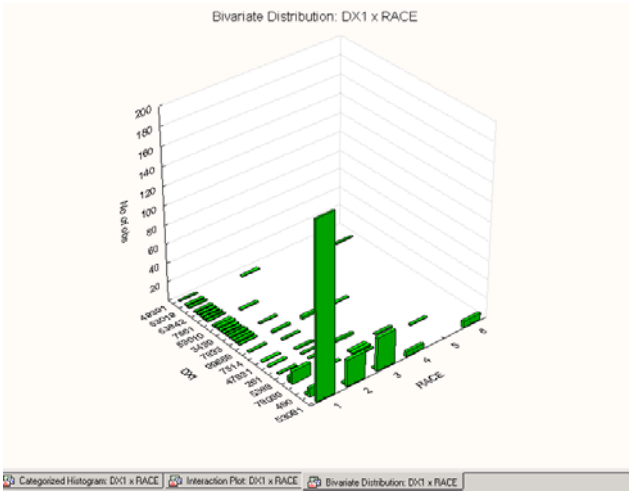

Figure 4. Graph of bivariate distribution For DX1 and Race.

2-Way Summary Table: Observed Frequencies

	WHITE	BLACK	HISPANIC	ASIAN OR PACIFIC ISLANDER	NATIVE AMERICAN	OTHER	TOTAL
53081	182	28	35	5	0	7	257
Row %	70.82%	10.89%	13.62%	1.95%	0.00%	2.72%	
5533	9	0	0	0	0	0	9
Row %	100.00%	0.00%	0.00%	0.00%	0.00%	0.00%	
53011	14	1	2	0	1	0	18
Row %	77.78%	5.56%	11.11%	0.00%	5.56%	0.00%	
9974	5	0	1	0	0	0	6
Row %	83.33%	0.00%	16.67%	0.00%	0.00%	0.00%	
5070	0	2	3	1	0	0	6
Row %	0.00%	33.33%	50.00%	16.67%	0.00%	0.00%	
Totals	235	39	46	7	1	8	336

Table 3. 2-way summary table for DX1 and Race

3.2. EM cluster analysis

The advanced EM clustering is a probability-based clustering technique. It adjusts the parameters in the model by iterating over an E-step (Expectation) and M-step (maximization). The data mart was mined with the same setting as in 3.1 using the EM clustering algorithm. The result shown that cluster 1 has a higher probability density and higher percentage of occurrence and the 1-13 days of the length of stay is also the norm. Table 4 shows the clustering result with the LOS.

Statistics for continuous variable: LOS
Number of clusters: 2
Total number of training cases: 544

	Cluster 1	Cluster 2	Overall
Minimum	1.00000	14.0000	1.0000
Maximum	13.00000	157.0000	157.0000
Mean	5.02078	30.6435	10.9489
Standard deviation	3.11386	24.8248	266.7735

Table 4 Summery of 2 clusters for LOS

Similarly the cluster 1 have focused on to conducting further analysis on the relationship between DX1 and Gender, and DX1 and Race using statistic technique of cross tabulation table and frequency tables.

Relationship between DX1 and Gender

Again, it has been found that there are 79.90% of patients having DX1 code, 53081, and female of these patients has slightly higher percentage than male (table 5 and figure 5). There is also no evidence shown that gender is a major influence.

2-Way Summary Table: Observed Frequencies

	MALE	FEMALE	Row
53081	187	147	334
Total %	44.74%	35.17%	79.90%
53011	12	8	20
Total %	2.87%	1.91%	4.78%
9974	5	4	9
Total %	1.20%	0.96%	2.15%
Totals	240	178	418
Total %	57.42%	42.58%	100.00%

Table 5. 2-way summary table for DX1 and Gender

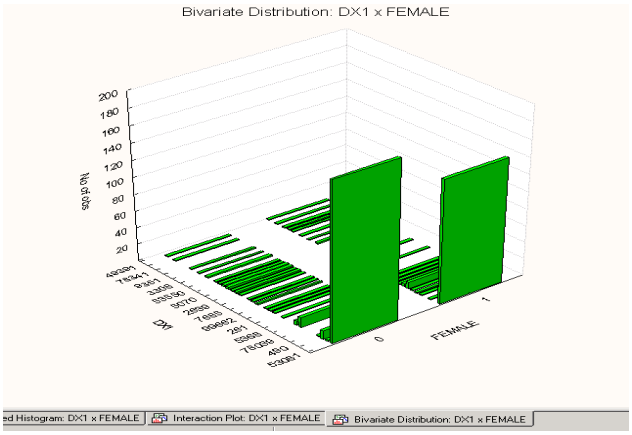

Figure 5. Graph of Bivariate Distribution For DX1 and Gender

Relationship between DX1 and Race

Similarly, the Race analysis (table 6) showed the normal distribution. Without the further analysis of the geometric distribution such as urban, rural data, it is difficulty to associate the DX1 group with the race.

2-Way Summary Table: Observed Frequencies

	WHITE	BLACK	HISPANIC	ASIAN OR PACIFIC ISLANDER	NATIVE AMERICAN	OTHER	Total
53081	177	27	32	5	0	7	248
Total %	56.01%	8.54%	10.13%	1.58%	0.00%	2.22%	78.48%
53011	14	1	2	0	1	0	18

Total %	4.43%	0.32%	0.63%	0.00%	0.32%	0.00%	5.70%
9974	5	0	1	0	0	0	6
Total %	1.58%	0.00%	0.32%	0.00%	0.00%	0.00%	1.90%
.							
Totals	229	32	40	6	1	8	316
Total %	72.47%	10.13%	12.66%	1.90%	0.32%	2.53%	100.00%

Table 6 2-way summary table for DX1 and Race.

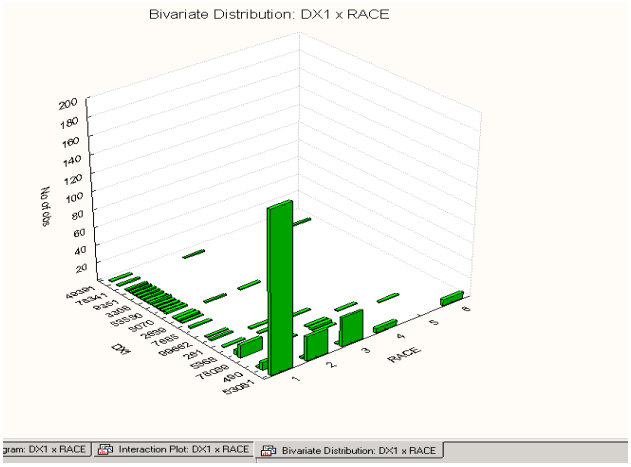

Figure 6 Graph of bivariate distribution For DX1 and Race

From both K-means cluster and EM cluster analyses, it has been found out that the patients who have DX1 code 53081 and are white are more likely having short time period staying in the hospital. Although it can't draw any conclusions for this analysis along, some hypothesizes might be built from it after considering other factors such as age, income, insurance, and region.

4 Clustering with different predictor variable

In this clustering analysis, the predictor variable has been changed from LOS to Age. The new distribution is shown in figure 7 where cluster 1 has the highest probability density and narrowest range of age than the others. Some meaningful information is listed in table 6. For example it can clearly see that more than 87% of patients of age 0 to age 3 have Medicaid and 56% of them are from lower household income families.

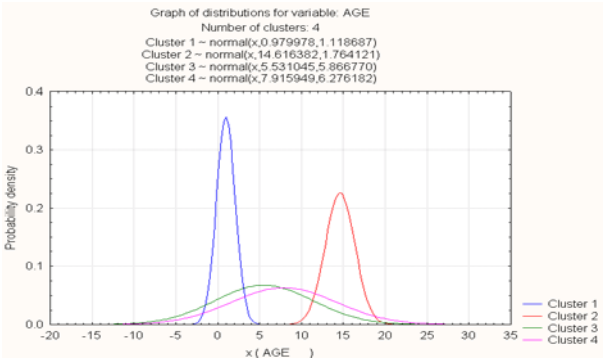

Figure 7. Graph of distributions for age

Statistics for continuous variable: AGE					
	C1	**C2**	**C3**	**C4**	**Overall**
Min	0.0000	12.000	0.0000	0.0000	0.00000
Max	3.0000	17.000	18.0000	18.000	18.00000
Mean	0.9799	14.616	5.53105	7.9159	5.31247
Sta dev	1.1188	1.7658	5.86724	6.2775	35.72623

Frequency table for categorical variable: PAY1					
Primary Payer Type	**C1**	**C2**	**C3**	**C4**	**Total**
(2) Medicaid	48	8	14	31	101
(3) Private including HMO	0	0	106	10	116
(4) Self-pay	2	0	0	1	3
(6) Other	5	2	0	2	9

Frequency table for categorical variable: ZIPINC					
Median Household Income	**C1**	**C2**	**C3**	**C4**	**Total**
(1) $1-$24,999	7	0	5	0	12
(2) $25,000-$34,999	24	6	28	3	61
(3) $35,000-$44,999	0	0	18	41	59
(4) $45,000 and above	24	4	69	0	97

Table 6. Information of age, primary payer, and household income for each cluster

Figure 8 shown the cluster 1 statistical information for the in-hospital death, gender of patient, length of stay, and race. For example, 2% of patients died during hospitalization; 56% of them are male; 74% of them stay in hospitals less than 20 days; 47% of them are white, 15% of them are black and 28% of them are Hispanic. There is one noticeable fact: while most of patients stay in hospitals less than 70 days, 2% of them do stay much longer. Further analysis can be conducted to profile these patients.

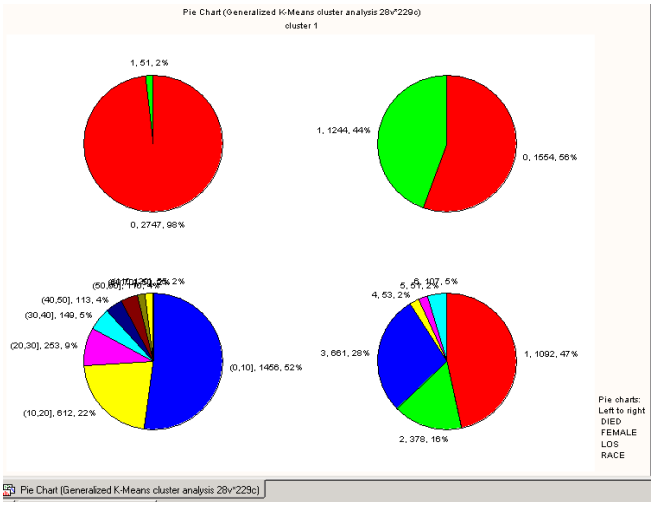

Figure 8. Pie chart of Died, Gender, Length of stay, and race for cluster 1.

From these clustering analyses, it also know that patients in cluster 1 are most likely either from East or West region. Two subgroups in the west region within the cluster showing that patients who stay in hospitals less than 40 days and have Medicaid as primary payer are likely having income of $25,000-$34,999 or $45,000 and above (interesting). This information can be used by the healthcare worker to better understand the GERD patient profiles and use them to develop a better treatment program.

5 Conclusions

Intelligent data analysis on the huge healthcare data sets is an important data mining applications. A limited healthcare data mining on GERD patients using the nationwide NIS data sets has been conducted. The results can be used by the healthcare worker to develop a better healthcare program for these patients nationally. Further studies can be conducted using comprehensive association rules and clustering algorithms to identify the meaningful information from the huge NIS database data.

Reference

[1] Overview of the Nationwide Inpatient Sample, http://www.hcup-us.ahrq.gov/nisoverview.jsp

[2] Overview of HCUP, http://www.hcup-us.ahrq.gov/overview.jsp

[3] Statistica Data Miner, StatSoft, 2002.

[4] Karen A. Wager, Frances Wickham Lee, John P. Glaser, and Lawton Robert Burns, Managing Health Care Information Systems: A Practical Approach for Health Care Executives, Wiley, 2005.

[5] Jiang B. Liu, Umadevi, and Daizhan Cheng, "A Distributed Knowledge Extraction Data Mining Algorithm", Proceedings of First International Symposium of Computational and Information Science, Shanghai, December 2004, pp. 769-774.

[6] Jiang B. Liu and Jun Han, "A Practical Knowledge Discovery Process for Distributed Data Mining," Proceedings of 11th International Conference on Intelligent Systems, Boston, MA, July 2002, pp 11-16.