Data Mining for Hidden Groups in Hospital Populations

Anthony M. Dymond, Dymond and Associates, Concord, CA

ABSTRACT

A data mining case study is presented illustrating exploration for hidden inpatient subpopulations. Some of the major characteristics of data mining and their relation to business goals are reviewed.

Six subpopulations are tentatively identified. Two subpopulations correspond to mental illness and alcohol/drug abuse. Four other subpopulations reflect medical and surgical diseases, segregated along lines of major diagnostic category, age, numbers of other diagnoses, and income.

Some suggestions are presented for using cluster analysis to find potential hidden groups, and for the interpretation and validation of these groups using visualization and discriminant analysis techniques.

INTRODUCTION

Data mining is an exploratory process aimed at discovering previously unknown features in a database. Although some of the information to be discovered may be hypothesized to be embedded in the database, it will usually take more than casual inspection to demonstrate its existence. Within a business context, the results of data mining should also be useful and understandable, in addition to being novel (Fayyad et al., 1996; Brachman and Anand, 1996).

Almost any visualization or analytic procedure can offer a valid approach for data mining (Elder and Pregibon, 1996). In addition to classical statistics, analytic techniques can also include induction (tree-based models) and neural nets. Each data mining project may require its own unique set of visualization and analytic techniques.

In addition to the various technical considerations, it is important to recognize that data mining frequently depends on continuous interaction between the data miner and subject matter experts who have both tangible knowledge as well as intuitive insights into the nature of the business process and the databases. These teams interact throughout the project, moving iteratively through a process of hypothesis formulation, data base exploration, and modeling and hypothesis testing. The SAS Institute Inc. refers to this as the SEMMA process (Sample, Explore, Modify, Model, and Assess) (SAS Communications, 1996). The underlying data bases are examined, and selected cases and variables are used for preliminary analyses. These results suggest initial hypotheses and the modified data samples needed to model and test them. Finally, the results are assessed for understandability and usefulness to the business goals.

Some of these process, such as data selection, cleaning, and transformation, resemble the construction of a data warehouse (Kimball, 1996; Mattison, 1996). Indeed, data mining is one of the “data exploitation” operations that motivate building a data warehouse. Almost all of the processes in data warehousing can be required for data mining. One general exception is that, since data mining is often a one-time activity, it does not usually involve data warehouse scheduling and retention concerns beyond synchronization of the particular data sets required for the analyses.

HOSPITAL INPATIENT SUBGROUPS

Previous work (Dymond, 1996) has tentatively identified three subpopulations within a VA hospital inpatient population when both inpatients and their related outpatient activities are considered. Two of these groups correspond to recognized populations of World War II (WWII) and Korea era veterans who mainly utilize traditional medical and surgical resources, and Vietnam era patients with a much higher tendency to be in mental and alcohol/drug related categories. A third subgroup may exist that is characterized mainly by a tendency to extensively utilize outpatient services. This third group of patients is distributed over all diagnostic categories.

The purpose of this data mining study is to probe for further substructure in just the inpatient databases. There are both clinical and business management issues motivating this project. Clearly, one patient management plan may not be effective if there are multiple distinct patient subgroups.

Beyond developing an inpatient management plan for a given hospital, there are issues about managing patients across a network of hospitals.

Hospitals in VA networks can range from large urban facilities with close relationships to local university teaching hospitals, to rural facilities with more limited programs or emphasis on specialty areas such as mental health. Questions now arise about the existence of similar subpopulations in different facilities, differences between the facilities themselves, and the necessity for management plans that consider characteristics of both the facilities and the patient subpopulations. In defining characteristics of the facilities, one can also hypothesize that characteristics of the local patient populations may be highly useful in characterizing a facility and in contrasting it with other facilities in the network.

Data for this study was available from inpatient discharges in 1995 from a network of eight VA hospitals. Data from all hospitals is rolled up to an IBM® mainframe running SAS® software. Some analyses were run on the mainframe, and some data was downloaded to a PC running SAS software.

This data was screened for outliers, and a variety of visualization and analytic procedures were carried out. Particular attention will be paid to the results of cluster analysis on a subset of data from one of the larger metropolitan hospitals, where several patient subgroups have been tentatively identified.

Cluster analysis is a technique of choice to search for unknown groups in a population. However, cluster analysis can be a challenging multivariate approach when clusters are marginally separated. The SAS/STAT® Users Guide (1990) provides a good discussion of the different types of cluster analysis and the use of reported measures, such as the cubic clustering criterion, in determining where valid clusters may exist.

In this study, a variety of clustering techniques were tried. Several of them, in particular Ward clustering, suggested six clusters might exist in the data. In order to explore this finding, the output of PROC CLUSTER was processed by PROC TREE with the option “nclusters=6” set to assign a cluster number between one and six to each record in the test data set. This allows the test data set to be subgrouped by cluster number and submitted to further visualization and analytic processes. Figure 1 shows the cluster analysis variables plotted by the cluster number.

The output of PROC TREE was then processed by PROC DISCRIM where the cluster numbers were used to identify the discriminant classes. PROC DISCRIM can be used to estimate the validity of groups by estimating the probability of misclassification into the groups. It can also produce canonical variables that can be plotted to provide some visual insight into the data structure. Table 1 shows the classification errors for this data.

Figure 1 can be examined to attempt to identify the candidate clusters and to gain an intuitive sense if they are reasonable. A tentative interpretation of these clusters is:

Cluster 1: Elderly patients (age>70) with hospital discharges included in the broader range of major diagnostic categories (MDCs 4-16 [see Appendix 1]). They have a higher than average number of diagnoses (approximately 5), but tend to be kept in only one bedsection during their hospital stay.

Cluster 2: Patients between the ages of 50 and 70 with MDCs concentrated in the core medical and surgical areas (MDCs 4-8).

Cluster 3: Patients with ages between 60 and 70 and with inpatient stays focused on vascular and heart disease (MDC=5). In addition to their primary circulatory disease, these patients also have other diagnosed diseases (between 4 and 7), and tend to be transferred between multiple bedsections (2 or 3) during their hospital stay.

Cluster 4: Patients around age 40 whose admission was associated with alcohol/drug abuse (MDC=20).

Cluster 5: These patients closely resemble those in cluster 2. They may have fewer respiratory and circulatory diseases, but their most distinguishing characteristic may be that they have the highest income of any patient cluster.

Cluster 6: Patients around age 40 whose admission was associated with mental illness (MDC=19). These patients also have by far the longest length of stay (>25 days) as inpatients.

Table 1 displays discriminant analysis classification errors between the clusters. For example, 80.0% of cluster 1 patients were correctly classified into cluster 1. Of the misclassified cluster 1 patients, most (11.8%) were misclassified into cluster 2. Similarly, cluster 2 misclassifications are generally placed into cluster 1. Clusters 3, 4, and 6 all classify with higher than 90% accuracy. Cluster 5 classifies with 80.3% accuracy, with most of the misclassifications placed into clusters 1-3.

Figure 1 and Table 1 taken together provide some insights into the interpretation and validity of the six clusters. Clusters Four (alcohol and drug) and Six (mental) represent well known populations. Cluster Three (circulatory) may be a reflection that there are more discharges into MDC five then into any other MDC. After extraction of the circulatory patients, Cluster One represents a broad range of mainly medical/surgical discharges of very old patients, and Cluster Two represents a smaller range of medical/surgical discharges affecting the remaining slightly younger patients. Based on just clinical criteria, Cluster Five should probably be merged with Cluster Two.

The clusters are consistent with clinical expectations. Classification accuracies are encouraging, and misclassifications are explainable based on variable range overlap between the clusters.

DISCUSSION

There appear to be five clusters in this inpatient population (after allowing Cluster five to merge into Cluster Two). There is a cluster for mental diseases and another for alcohol/drug related disorders. A third cluster encompasses a variety of circulatory diseases. The remaining two clusters account for other medical/surgical diagnoses, with one cluster focusing on elderly patients with a broad range of diagnoses, and the final cluster focusing of younger patients with a narrower range of diagnoses.

Collecting multiple variables into a smaller set of discharge clusters can help to simplify the analyses and to discover hidden structures in the data. Five discharge clusters may summarize much of the information associated with a given discharge. The five clusters may also indicate areas to consider for future patient management plans. These clusters may prove to be valuable tools to use in further analyses such as comparing hospitals based on their discharge characteristics.

Arriving at a proposed set of clusters is an iterative process. Some issues to consider are:

- Carefully screen and select both variables and cases. Transform variables as needed
- Try several different clustering techniques.
- Evaluate results for different numbers of clusters.
- Add and remove variables and repeat the analyses.
- Select several random data samples and compare the analyses results.
- If large clusters are found, remove the cases and repeat the analyses.
- Utilize data visualization and discriminant classification to help interpret and validate the results.
- Utilize subject matter experts.

Choosing variables for cluster analysis can be challenging. For example, Cluster 5 may have separated largely due to differences on the income variable. This variable was originally included because of a hypothesized relation between income and both clinical diagnoses and disease severity. However, the results now suggest repeating clustering without the income variable.

Another approach to evaluating variables comes from the canonical variables produced from discriminate analysis. These canonical variables can be rotated and used to estimate the relative contribution of individual variables to group separation.

Ordinal variables are admissible only to the extent that they are believed to mimic well behaved continuous variables. Major diagnostic category is an ordinal variable that has contributed to this study. Alternate approaches to evaluate the interactions and contributions of categorical variables include logistic and loglinear analysis.

CONCLUSION

This project illustrates many of the characteristics of data mining. The process is business driven, and involves a complex iteration involving the data miner, subject matter experts, data selection, and hypothesis testing utilizing an assortment of data visualization and analytic tools.

Good progress has been made towards the data mining goals of finding novel, useful, and understandable information. It is important for the data miner to maintain this business focus, and to value the products in terms of how they affect business and clinical plans and outcomes. Even at an intermediate stage, this study is successful to the extent that it can begin to provide management with a sense of possible directions for future planning.

From a technical perspective, this project illustrates some of the issues encountered when using cluster analysis to search for subgroups, and offers some pragmatic insights into how to approach these situations.

APPENDIX 1
Major Diagnostic Categories

1 - Nervous System
2 - Eye
3 - Ear, Nose & Throat
4 - Respiratory
5 - Circulatory
6 - Digestive
7 - Liver & Pancreas
8 - Muscle, Bone, & Connective
9 - Skin, Subcutaneous, & Breast
10 - Endocrine & Metabolic
11 - Kidney & Urinary
12 - Male Reproductive
13 - Female Reproductive
14 - Pregnancy
15 - Newborn
16 - Blood & Related
17 - Myeloproliferative
18 - Infectious & Parasit.
19 - Mental
20 - Alcohol/Drugs
21 - Injuries & Toxic
22 - Burns
23 - Health Visit
24 - Multi. Sig. Trauma
25 - HIV Infections

REFERENCES

Brachman, R.J. and Anand, T., “The Process of Knowledge Discovery in Databases,” In: Fayyad, U.M., Piatetsky-Shapiro, G., Smyth, P., and Uthurusamy, R., eds. (1996), *Advances in Knowledge Discovery and Data Mining,* Cambridge: MIT Press.

Dymond, A.M., (1996), “Preliminary Data Mining for Subgroups in a Hospital Inpatient Population,” Proceedings of the Fourth Annual Western Users of SAS Software Regional Users Group Conference, San Francisco, California.

Elder, J.F. and Pregibon, D., “A Statistical Perspective on Knowledge Discovery in Databases,” In: Fayyad, U.M., Piatetsky-Shapiro, G., Smyth, P., and Uthurusamy, R., eds. (1996), *Advances in Knowledge Discovery and Data Mining,* Cambridge: MIT Press.

Fayyad, U.M., Piatetsky-Shapiro, G., Smyth, P., “From Data Mining to Knowledge Discovery: An Overview,” In: Fayyad, U.M., Piatetsky-Shapiro, G., Smyth, P., and Uthurusamy, R., eds. (1996), *Advances in Knowledge Discovery and Data Mining,* Cambridge: MIT Press.

Kimball, R., (1996) *The Data Warehouse Toolkit*, New York: John Wiley & Sons.

Mattison, R., (1996*), Data Warehousing: Strategies, Technologies, and Techniques*, New York: McGraw-Hill.

“Data Mining Reveals the Diamonds in Your Databases,” (1996*), SAS Communications*, 22, 15-20.

SAS Institute Inc., (1990), *SAS/STAT User’s Guide, Version 6, Fourth Edition*, Cary, NC: SAS Institute Inc.

ACKNOWLEDGMENTS

SAS and SAS/STAT are registered trademarks or trademarks of SAS Institute Inc. in the USA and other countries. IBM is a registered trademark or trademark of International Business Machine Corporation. ® indicates USA registration.

Other brand and product names are registered trademarks or trademarks of their respective companies.

CONTACT INFORMATION

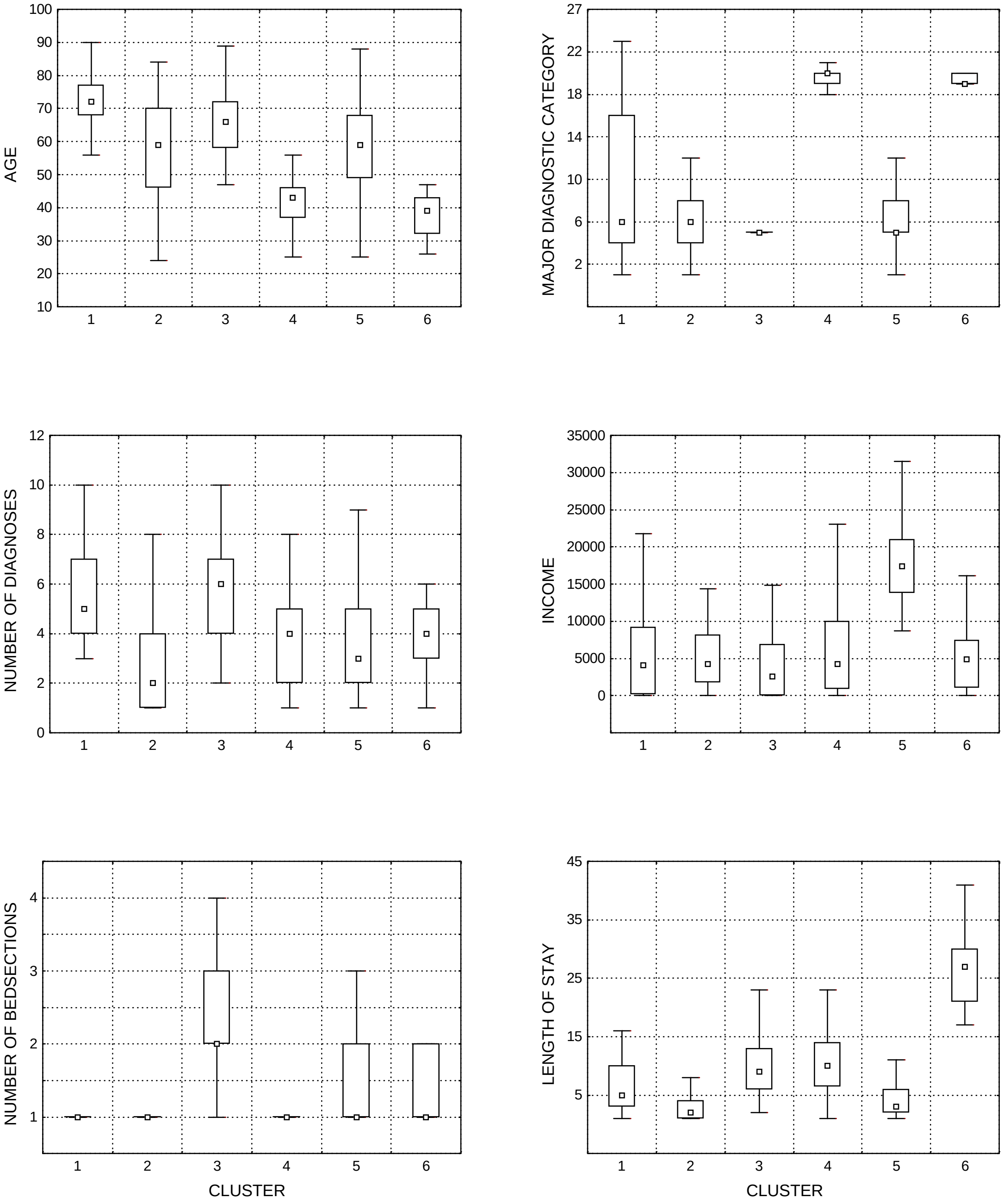
Anthony M. Dymond, Ph.D.
Dymond and Associates
4417 Catalpa Ct.
Concord, California 94521

(510) 798-0129
dymond@ccnet.com

TABLE 1
CROSSVALIDATION CLASSIFICATION RESULTS FOR SIX CLUSTERS

	Number of Observations and Percent Classified into Cluster						
From Cluster	1	2	3	4	5	6	Total
1	88 80.0	13 11.8	0 0.0	5 4.5	4 3.6	0 0.0	110 100.0
2	11 10.6	84 80.8	0 0.0	2 1.9	7 6.7	0 0.0	104 100.0
3	1 1.5	0 0.0	63 95.5	0 0.0	0 0.0	2 3.0	66 100.0
4	2 2.6	0 0.0	0 0.0	69 90.8	2 2.6	3 3.9	76 100.0
5	4 5.6	6 8.5	4 5.6	0 0.0	57 80.3	0 0.0	71 100.0
6	0 0.0	0 0.0	0 0.0	1 4.3	0 0.0	22 95.7	23 100.0
Total Percent	106 23.6	103 22.9	67 14.9	77 17.1	70 15.6	27 6.0	450 100.0

FIGURE 1
Box and Whisker Plots of Variables by Cluster


AGE
MAJOR DIAGNOSTIC CATEGORY
NUMBER OF DIAGNOSES
INCOME
NUMBER OF BEDSECTIONS
CLUSTER
LENGTH OF STAY
CLUSTER