Association for Information Systems
AIS Electronic Library (AISeL)

AMCIS 1998 Proceedings

Americas Conference on Information Systems (AMCIS)

December 1998

Identifying and Modeling Relationships Between Diagnostic and Procedure Codes

William Spangler
West Virginia University

Jerrold May
University of Pittsburgh

David Strom
University of Arkansas Medical Center University of Pittsburgh

Identifying and Modeling Relationships Between Diagnostic and Procedure Codes

William E. Spangler
College of Business and Economics
West Virginia University

Jerrold H. May
Joseph M. Katz Graduate School of Business
University of Pittsburgh

David P. Strum
University of Arkansas Medical Center

Luis G. Vargas
Joseph M. Katz Graduate School of Business
University of Pittsburgh

Abstract

This research explores the relationships between diagnostic and procedural codes in a medical setting, with the objective of developing a general classification and predictive model. We describe the inherent relationship between codes in the context of a general data model, but note that the model is somewhat tenuous and requires extensions through other data analytic / data mining techniques. One of those techniques is decision tree induction, which is described briefly as a possible supplement to the initial code-to-code patterns. The paper concludes with implications for future research, including the investigation of additional analytic techniques and the extension of the model into other domains where problem-solving has been codified.

Introduction

This paper describes induced relationships between problem and solution taxonomies in medical informatics, and indicates how decision tree induction, a popular data mining technique, can be used to enhance descriptions of those relationships. In the medical domain, the dominant problem taxonomy is the International Classification of Diseases (ICD-9) coding system, which indicates a patient's disease or condition. The corresponding solution taxonomy is the Common Procedural Terminology (CPT) system, which indicates the procedures performed in order to correct the problem. Because these codes are related to varying degrees and in various ways, we are seeking to understand and describe the nature of the relationships, with the hope of constructing a general problem-solution model applicable across domains. The overall research goal is to understand how, and to what extent, hierarchical knowledge structures pertaining to problems and solutions might be related.

This research initially has two major objectives, which are:

1. to discover useful patterns in surgical patient data, including the relationships between ICD-9 and CPT codes, and to ascertain the impact of patient demographic and other case data on those relationships
2. to determine whether an integrated taxonomy of diagnostic and procedure codes might be constructed, and if so, how it might be structured.

Linking problems and solutions has important implications for cross-domain activities such as scheduling and auditing. Scheduling is facilitated through the ability to use induced patterns to predict procedures from diagnostic assessments and other case data. Conversely, given an observed procedure and other information about a patient, auditing seeks to determine what types of diagnoses would be expected, and how they compare to actual observations. Both of these activities are dependent on the various parameters that comprise each case, and on how those parameters combine in certain ways to link case problems with solutions. When a particular case is under consideration, the linkages provide guidance to a decision maker in determining, and then searching for, the parameters that will be most useful in making a decision.

Preliminary Induction of Data Patterns Using Relative Frequencies

Collected over six years, the empirical data include 59864 separate cases (patient surgeries), each containing 23 attributes detailing the diagnoses, procedures, patient demographic information, and information about the individual surgeon and anesthesiologist. Initial patterns were derived simply by compiling and ranking the relative frequencies of procedure codes for each diagnostic code (and vice versa). At this point, no other information was used to characterize the patterns. For example, Figure 1(a) illustrates the most frequent procedure codes related to ICD-9 180.9 (malignant neoplasm of cervix).

CPT	Count	Pct	Cumul Pct
58210	**153**	**20.99**	**20.99**
77762	102	13.99	34.98
77761	43	5.90	40.88
58150	38	5.21	46.09
58200	38	5.21	51.30
57520	24	3.29	54.60
57500	14	1.92	56.52
…	…	…	…

(a) -- ICD-9 = 180.9

ICD-9	Count	Pct	Cumul Pct
180.9	**168**	**74.34**	**74.34**
182.0	32	14.16	88.50
183.0	7	3.10	91.59
789.3	7	3.10	94.69
…	…	…	…

(b) -- CPT = 58210

Figure 1. One-to-many Relationship Bewteen ICD-9 180.9 and CPT 58210

Simple frequency analysis of the data indicates that the relationships between ICD-9s and CPTs can be loosely characterized in data modeling terms: i.e., as one-to-one, one-to-many, many-to-one, and many-to-many. For example, the relationship between the ICD-9 in Figure 1(a) and its associated CPTs can be characterized initially as a one-to-many relationship. There are a number of potential CPTs for this one ICD-9, with no single CPT dominating (CPT 58210 is the most frequent, accounting for about 21% of the cases). Conversely, many CPTs are linked primarily back to this single ICD-9. Figure 1(b), for example, shows that although CPT 58210 is linked to a number of ICD-9 codes, its dominant linkage is with ICD-9 180.9. That is, if CPT 58210 is observed, it is highly likely (74.3 %) that the original diagnostic code was 180.9. This is true for other CPTs as well. Although not shown here, the frequency for CPT 77762 is even higher (82%).

Unfortunately, due to the inherently messy nature of real world data, the situation is not this simple. While it might appear from the preliminary analysis that this is generally a one-to-many relationship (with some minor exceptions), the tidy nature of the relationships begins to break down upon inspection of subsequent CPTs. For example, only 27% of cases associated with CPT 58200 can be linked back to ICD-9 180.9, and the situation is as bad, or worse, for many other CPTs. In the context of fuzzy sets, the exceptions to the general data model suggest that there are varying *degrees of membership* in the set of CPTs belonging to the single ICD-9. That is, CPT 58210 has a 74% degree of membership in the set of CPTs attached to ICD-9 180.9, whereas CPT 58200 has only a 27% degree of membership in the same set -- *given what we currently know about the relationship*.

To summarize the analysis thus far:

1. a characterization of ICD-9/CPTrelationships in simple data modeling terms using only the codes themselves is premature, and
2. the major and minor exceptions to the general data model suggest that more information is needed in order to more accurately identify the relationships between codes
3. "more accurately identifying the relationships" likely entails evolving the model of the relationship patterns, with the goal of increasing the degrees of set membership observed in the preliminary analysis

Enhancing The Initial Analysis Using Decision Tree Induction — A Brief Example

The use of decision tree induction (Quinlan, 1993) is motivated by the notion that the knowledge of simple ICD-9/CPT relationships can be supplemented with additional semantic linkages reflecting knowledge of other case attributes. The goal is to more precisely define the relationships between codes, and by doing so to improve the accuracy of predictions obtained from the relationships.

As shown in Figure 2, an induced decision tree improves the relatively tenuous link between CPT 58200 and ICD-9 180.9. It does so by introducing additional information from the patient record, in this case other procedures performed on the patient and the patient's age. That is, given that one of a specific set of procedures (i.e., 38562, 57520, or 58960) was performed, and given that the patient was under 45 years old, there is a 61.7% chance that the patient's diagnostic code was 180.9 (as compared to 27% in the simple code-to-code comparison). This suggests that by introducing additional information, decision tree induction allows a more accurate characterization of the relationship.

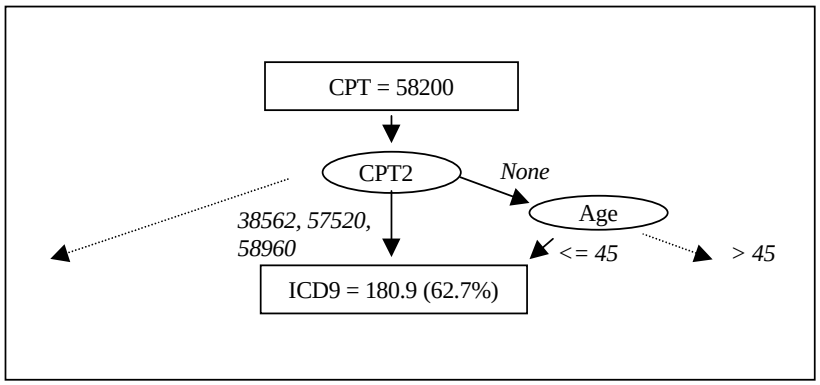

Figure 2. A Decision Tree Linking CPT with ICD-9 via Case Attributes

Conclusions and Future Research

Initial findings indicate that the additional semantic information provided by decision tree induction enriches the inherent relationship existing between ICD-9 and CPT codes. By making explicit the semantic linkages inherent in the taxonomy, the various exceptions in the general data models can be better understood and explained. As an aside, it is important to note that the exceptions to the general data model appear to vary dramatically. In some relationships, it is difficult to ascertain *any* type of general data model, suggesting that other methods such as decision tree induction would replace, rather than supplement, the simple code-to-code relationships. In other cases, the general model is quite accurate. For example, ischemic heart disease, a form of heart failure, can be modeled -- with relatively few exceptions -- as a one-to-many relationship with several heart bypass procedures. In this case, the need to supplement the data model with additional analysis is comparatively less.

This work is evolving toward a more general model that would incorporate a number of methods for modeling patterns in code relationships. Alternative and complementary methods would include the use of case-based reasoning, neural networks, and traditional statistical techniques such as discriminant analysis. Neural networks and discriminant analysis, as well as the rule induction technique shown here, are most useful when a reasonable number of general patterns can be identified, and when most historical and future cases can be characterized within the context of those patterns. Case-based reasoning, on the other hand, is useful for the identifying and characterizing the outlier cases -- i.e., those cases that might appear rarely in the data set but which nevertheless become useful when the particular sets of parameters shown in those cases are encountered (Kolodner, 1993). Because there is evidence of both general patterns and outliers in our data set, a multiple technique approach seems appropriate for future research.

In addition, a number of other domains, including process control environments, exhibit problems and prescribed solutions that are formalized in various ways (Pople et al, 1994). This research likely would be applicable in those areas as well.

References

Kolodner, J. (1993). *Case-based Reasoning*. San Mateo, CA: Morgan Kaufmann.

Pople, H. E., Spangler, W. E. and Pople, M. T. (1994). *EAGOL: An Artificial Intelligence System for Process Monitoring, Situation Assessment, and Response Planning*. Tenth IEEE Conference on Artificial Intelligence for Applications, San Antonio, TX.

Quinlan, J. R. (1993). *C4.5 : Programs for Machine Learning*. San Mateo, CA: Morgan Kaufmann.