Association for Information Systems
AIS Electronic Library (AISeL)

AMCIS 2000 Proceedings

Americas Conference on Information Systems (AMCIS)

2000

The Impact of Data Characteristics on the Selection of Data Mining Methods for Predictive Classification

William E. Spangler
Duquesne University, spangler@duq.edu

Jerold H. May
University of Pittsburgh, jerrymay@katz.pitt.edu

David P. Strum
Queens University, strumd@post.queensu.ca

Luis G. Vargas
University of Pittsburgh, vargas@katz.pitt.edu

The Impact of Data Characteristics on the Selection of Data Mining Methods for Predictive Classification

William E. Spangler, A.J. Palumbo School of Business Administration, Duquesne University, spangler@duq.edu
Jerrold H. May, Joseph M. Katz Graduate School of Business, University of Pittsburgh, jerrymay@katz.pitt.edu
David P. Strum, Department of Anesthesiology, Queens University, strumd@post.queensu.ca
Luis G. Vargas, Joseph M. Katz Graduate School of Business, University of Pittsburgh, vargas@katz.pitt.edu

Abstract

This research-in-process is exploring a contingency approach to the construction and selection of data mining models for predictive classification. This approach considers the structure of the data set and the relationships between and among the various attributes characterizing the data set, with the goal of selecting a model that provides greater insight into the data – and therefore predicts most accurately -- given a particular data structure. Preliminary results obtained from analysis of hospital patient records indicate that concentration indices, commonly used to measure firm concentration within an industry, are useful in characterizing data set structures and therefore in guiding the model selection process. The eventual goal of this research is the construction of a decision support system that can aid decision makers in the model selection task.

Introduction

This paper describes research-in-process into the selection of data mining models for predictive classification, with the eventual goal of building a decision support system to aid in model selection. Predictive classification is the process of 'guessing' into which category a new data item will fall, and then subsequently determining the true category of the item and evaluating the accuracy of the original guess. Based on problem characteristics in situations when a subset of relevant information is available before the fact, this research seeks to discover (1) when one should build a predictive model -- i.e., when a model will beat the naive approach of simply guessing the most frequent category in a data set, and (2) how to choose the best modeling method assuming model construction is worthwhile.

The essential goal of this research is to improve predictive classification by choosing a model based on the structure of the data set under analysis. A number of data sets can be characterized generally as relationships between *problem* – or case -- descriptions, and *solutions* implemented to address or solve the problem. In the medical domain under study here, 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. Intervening in the relationship between problems and solutions are *situation variables*, which contain patient demographic (*age*, *sex*, etc.) and other information (*surgeon*, *anesthesiologist*, procedure times, etc). Thus, the mapping from problem to solution in this domain is as follows: From a case description of a patient's disease(s) (i.e., the ICD-9), *as well as* other case-related (situational) information, predict which procedure(s) is/are most appropriate for the case at hand.

Discovering patterns through concentration indices

We begin to characterize the relationships between diagnostic and procedure codes by examining the frequency distribution patterns of procedures within each diagnostic category. In order to compare the various distribution patterns across potentially hundreds of different diagnostic categories, we quantify the patterns in a single number called a *concentration index (CI)*. A concentration index is a concept borrowed from the field of industrial economics, which attempts to characterize the concentration (and dominance) of various firms within a particular industry (Shepherd, 1997). Because we are attempting to characterize the concentration of procedure codes within a particular diagnostic code, the use of concentration indices in the ostensibly unrelated field of data mining seems appropriate.

The degree of concentration can be measured using various methods, which collectively tend to produce

similar – although not identical – results. We are focusing on four of the methods, as follows:

- Hirschman-Herfindahl index
- 'N-firm' concentration ratio
- Rosenbluth index
- Entropy index

The Hirschman-Herfindahl (HH) index serves as an example of these types of methods. The HH method

$$\sum_{i=1}^{n} p_i^2$$

calculates an index from the following formula:

Where:

n = number of firms
P_i is the percentage share of the ith firm (* 100) -- ($i = 1 \ldots n$).

The output of the formula is a number between 0 and 10,000. Specifically, larger numbers indicate more highly concentrated industries (where one or a few firms dominate), while smaller numbers indicate more diverse industries (where the market is shared somewhat equally by a number of firms).

In the domain of diagnostic and procedure codes, the industry concept is replaced by a diagnostic code, while the firms are replaced by procedure codes assigned to the diagnosis. Thus, each diagnostic code (industry) will have associated procedure codes (firms), and the codes will be variously concentrated, as calculated by any of the four methods listed above.

Data Mining Models

Given a means of characterizing problem-solution relationship patterns using concentration indices, the next step is the selection of data mining models for inclusion in this study. We adopted an MIS end-user perspective in selecting the models, in that we considered only widely accepted models with wide commercial acceptance. That is, included models should be readily available to a decision maker, and professionally-implemented and maintained. With those constraints, we focus initially on three basic models and implementations:

- Decision tree induction (Implementation: *See5, CART* – see (Breiman, et al., 1984; Quinlan, 1993))
- Neural networks (Implementation: *SPSS/Clementine*)
- Linear Discriminant Analysis (Implementation: *Statgraphics*)

Data analysis

The basic research question entails the relationship between concentration indices and data mining models, and asks whether the concentration index for a particular diagnostic code differentially impacts the performance of data mining model candidates. Or, from an end-user's perspective, is the selection of a data mining model contingent on the structure of the data set – as measured by a concentration index? This general question in turn can be decomposed into a number of testable sub-questions, as follows:

1. Does the degree of concentration impact the accuracy of a data mining method? If so, in what way?
2. Does the impact vary across methods?
3. Is method selection (*or not*) contingent on the degree of concentration?

Based on these research questions, two areas of exploration seem feasible:

1. *Within methods* (Question 1) -- How does a single method perform (i.e., predict CPTs) as the concentration index moves from high to low (i.e., concentrated to dispersed)?
2. *Across methods* (Questions 2 & 3) -- How do the various methods compare to each other, based on different concentration indices? This strikes to the heart of the methods selection process.

We have begun to explore these questions through the analysis of patient data collected from three hospitals over a seven year period. The data include 59864 separate cases (patient surgeries), each containing 23 attributes detailing the diagnoses (ICD9s), procedures (CPTs), patient demographic information, and information about the individual surgeon, anesthesiologist, type of anesthesia, and various procedure times. From the entire patient data set, we selected patient records with ICD9 codes having at least 50 associated records (193 different ICD9s met this requirement, totaling 33385 records).

Initial analysis of the data has occurred in three stages. The first stage entailed calculating concentration indices for each of the 193 ICD9s using each of the four industry concentration formulas. With the HH index as the CI of interest, we sorted the ICD9s based on the CI, and using rules-of-thumb from industrial economics, classified each of the ICD9s as having either *high*, *medium* or *low* degrees of concentration (Shepherd, 1997). The second stage entailed determining the performance of each the four data mining implementations listed above – i.e., Decision Tree/See5(DT/S), Decision Tree/CART (DT/C), Neural Network/Clementine(NN/C), and Linear Discriminant Analysis/Statgraphics (LDA/S) – as well as

the performance of the naïve model (NM) of simply choosing the most frequently occurring CPT in each data set (which should work reasonably well when the ICD9 records are heavily concentrated in a single CPT). In this pre-validation stage, performance was derived by running each of the models against each of the 193 data sets, and calculating the percentage of correct classifications – i.e., derived CPT vs. actual CPT -- in each run. The third stage entailed performing a two-way ANOVA on the set of summarized ICD9 CI and model performance data. The ANOVA essentially determined whether performance (*accuracy of classification*) is a function of the interaction between *degree of concentration* (high, medium, low) and *model implementation* (DT/S, DT/C, NN/C, LDA/S, NM).

Preliminary Results

Although we have yet to validate the results obtained using a test sample and/or cross validation, initial results are intriguing. Our results can be summarized in the context of the research questions posed above.

Within Methods

- The performance of certain methods degrades linearly as the concentration decreases, including NM, NN/C and DT/C. Furthermore, the rate of degradation -- i.e., the slope of the line -- varies across methods. NN/C and NM degrade most rapidly, while DT/C degrades least rapidly.
- The performance of other methods cannot be characterized as linear. Notably DT/S initially degrades as concentration decreases, but then improves at lower levels of concentration.

Across Methods

The across methods analysis produced results corresponding to each of the three concentration levels. That is:

- When the concentration index of a particular ICD9 is *high* (i.e., dominated by one or a few CPTs), the DT/S method performs best
- When the concentration index of a particular ICD9 is *medium*, the NN/C and DT/S methods perform best
- When the concentration index of a particular ICD9 is *low*, the neural network method performs best

Continuing Research

In the short term, continuing research, primarily related to validation, is required in order to answer these questions definitively. The approach to validation entails choosing a validation strategy (or strategies), and then applying those strategies against each of the data sets using each of the data mining models. Candidate strategies would include the use of hold-out samples as well as various forms of cross validation. In addition to the simple notion of validating the preliminary results, other issues potentially could arise from the validation process. For example, from an end-user perspective, the ease of implementing different validation strategies will vary across methods. That is, some model implementations are more amenable to certain types of validation (e.g., n-fold cross validation) than are others.

References

Breiman, L., Friedman, J.H., Olshen, R.A. and Stone, C.J. *Classification and Regression Trees*, Wadsworth and Brooks, Monterrey, CA, 1984.

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

Shepherd, W.G. *The Economics of Industrial Organization*, Prentice Hall, Upper Saddle River, NJ, 1997.