December 2003

A Multistrategy Data Mining Approach to Classification

Mordechai Gal-Or
Duquesne University

William Spangler
Duquesne University

Jerrold May
University of Pittsburgh

A MULTISTRATEGY DATA MINING APPROACH TO CLASSIFICATION

Mordechai Gal-Or
Duquesne University
galor@duq.edu

William E. Spangler
Duquesne University
spangler@duq.edu

Jerrold H. May
University of Pittsburgh
jerrymay@katz.pitt.edu

Abstract

Our research explores the use of ensemble, *or* multistrategy learning *techniques for inducing and managing patterns of knowledge from organizational data. Specifically, we are exploring the use of data mining techniques in building an ensemble classification system – i.e., a system that incorporates multiple machine learning techniques to generate multiple models from existing data and make predictions about new observations. Our research is inspired and motivated by a real-world business problem. The emergence of the digital personal video recorder (PVR) is expected, over time, to cause profound changes in television viewing, as viewers use the new technology to time-shift viewing and skim over or eliminate 'in stream' commercials. This trend is a significant threat to television advertisers and service providers, because it jeopardizes the traditional means by which advertising finances so-called 'free' programming.*

Although a number of modeling methods are potentially useful for the analysis of television viewing data and the classification of specific viewer types, because of the complexity of the domain we cannot know a priori *which methods will be most accurate in specific situations. The effectiveness of a particular method is dependent on a number of factors, including the characteristics of the viewer, the prevalence of target viewers in the overall population, the specific viewer attributes to be predicted, asymmetry of misclassification costs, and other characteristics of the viewing data – including types of programs viewed, time of day, and so on.*

Because it is unlikely that any single method could perform optimally under these circumstances, we are developing an ensemble classifier composed of a number of different analytic methods. This classifier would process various television viewing data sets against each of the methods, and attempt to construct a single prediction about the viewer from the collective predictions of the various methods. We have conducted preliminary analyses of viewer data obtained from Nielsen Media Services, Inc. (NMSI), and developed an initial prototype of the data mining component from those analyses. Our initial study of viewing behavior for five target gender/age segments suggests that gains in performance are possible even with simple democratic voting schemes – i.e., where each method has a single vote. Our goal now is to determine whether we can do better by using more sophisticated combination strategies. We intend to approach the problem in two phases. The first phase will explore the combination of multiple methods in a controlled experiment using simulated *data, while the second will apply lessons learned from the controlled experiment to the analysis of actual television viewing data obtained from NMSI.*