Classification Trees as Proxies

Anthony Scime, Department of Computer Science, The College at Brockport, State University of New York, Brockport, NY, USA

Nilay Saiya, Department of Political Science, The College at Brockport, State University of New York, Brockport, NY, USA

Gregg R. Murray, Department of Political Science, Texas Tech University, Lubbock, TX, USA

Steven J. Jurek, Department of Political Science, The College at Brockport, State University of New York, Brockport, NY, USA

ABSTRACT

In data analysis, when data are unattainable, it is common to select a closely related attribute as a proxy. But sometimes substitution of one attribute for another is not sufficient to satisfy the needs of the analysis. In these cases, a classification model based on one dataset can be investigated as a possible proxy for another closely related domain's dataset. If the model's structure is sufficient to classify data from the related domain, the model can be used as a proxy tree. Such a proxy tree also provides an alternative characterization of the related domain. Just as important, if the original model does not successfully classify the related domain data the domains are not as closely related as believed. This paper presents a methodology for evaluating datasets as proxies along with three cases that demonstrate the methodology and the three types of results.

Keywords: Classification, Data Analysis, Data Mining, Proxy, Social Science

1. INTRODUCTION

One of the goals of data analysis is to find factors that characterize events and situations in the world. By understanding a current situation, policy makers, decision makers, and researchers can take actions directed toward changing society and individual lives for the better.

Sometimes decision makers cannot obtain all the data necessary to make a sound decision. For instance, a particular data point may be too costly, or it may be unobservable. In these cases, a proxy variable or attribute may be used. A proxy attribute is an attribute used in place of the unacquirable data. While not a direct measure of the desired data point, a good proxy attribute should be strongly related to the unobserved attribute of interest (Clinton, 2004). Proxies can introduce error in measuring the outcome (Kimball, Sahm, & Shapiro, 2008) but are necessary because the desired value is needed although unattainable. Sometimes multiple proxies exist and using a combination of these proxies can reduce proxy-introduced error (Lubotsky & Wittenberg, 2006; Trickett, Persky, & Espino, 2009). The extent of error

DOI: 10.4018/IJBAN.2015040103

is difficult or impossible to measure because the baseline, unattainable attribute is, well, unattainable.

At other times decision makers may want to assess the similarities between datasets. Here the decision maker hopes to make the decision using one dataset as a proxy for another dataset. Braslow and Humez (2014) and Hargittai (2005) investigated using survey data as a proxy for observation data. Observation data are more difficult and expensive to collect. Saunders, Bex, and Woods (2013) investigated the use of crowdsourcing data as a proxy for lab collected data in the medical domain. Crowdsourcing is well established in medical research for assembling large normative datasets. Of course, using one dataset as a proxy for another can also result in errors.

The ability to compare datasets could conceivably have great utility and real-world ramifications. One might want to know, for example, if gender or ethnic differences mattered in terms of election outcomes in one year but not in another, or if systemic differences like the institutional structure of a regime could correspond with the level of freedom in a state. Those studying the causes of war might be interested in whether or not the causes of civil and international war are similar and compare datasets on each to analyze the question. Comparing datasets on the causes of religious and secular terrorism would indicate if the determinants of both kinds of terrorism are the same. Analyzing a question in this way and showing differences between similar domains can carry ramifications for policy makers and researchers seeking to address the root causes of particular types of problems.

When a proxy attribute is used, regression analysis and other statistical techniques can test hypotheses to evaluate data against an expected outcome. These techniques inform a researcher about relationships between independent attributes including the proxy attributes and the dependent attribute or class attribute. This analysis can be used to study how an attribute influences an outcome while accounting for the other attributes that also influence the outcome. However, when attempting to substitute one dataset for another, regression and other statistical techniques may not provide sufficient information. Other techniques can be more insightful and practical than regression when predicting the interaction of attributes on the dependent attribute or class attribute (Andoh-Baidoo & Osei-Bryson, 2007; Chang, 2006). Classification can be used as an analysis technique when proxy attributes are used and the classification tree itself may act as a proxy tree for a similar domain.

This paper offers a methodology for evaluating the use of a dataset's classification model as a proxy model for a similar dataset; it then presents three cases that demonstrate the methodology and the three types of results. In this endeavor, the next sections describe classification analysis and its use as a mechanism for identifying proxy models. Then, it presents the three case studies regarding executive leadership, voter turnout, and terrorism followed by a conclusion.

2. CLASSIFICATION TREE CONSTRUCTION

Classification analysis constructs a classification tree or model, which is both explanatory and predictive. It is an inductive analysis of data, which finds internal patterns and relationships based on attributes and their values. Classification is used not only to predict the outcome of a future event but also to provide knowledge about the structure and interrelationships among the data (Osei-Bryson, 2004).

The C4.5 classification algorithm constructs models by looking at the past performance of input attributes (i.e., independent variables) with respect to a class attribute (i.e., a dependent variable) (Quinlan, 1993; Witten & Frank, 2005). The model is constructed inductively from records with known values for the class attribute. Attributes are sequentially selected from the dataset to construct the classification model using a divide-and-conquer algorithm that is driven by an evaluation