A TOPSIS Data Mining Demonstration and Application to Credit Scoring

Desheng Wu, University of Toronto, Canada
David L. Olson, University of Nebraska, USA

ABSTRACT

The technique for order preference by similarity to ideal solution (TOPSIS) is a technique that can consider any number of measures when seeking to identify solutions close to an ideal and far from a nadir solution. TOPSIS traditionally has been applied in multiple criteria decision analysis. In this article, we propose an approach to develop a TOPSIS classifier. We demonstrate its use in credit scoring, providing a way to deal with large sets of data using machine learning. Data sets often contain many potential explanatory variables, some preferably minimized, some preferably maximized. Results are favorable by a comparison with traditional data mining techniques of decision trees. Proposed models are validated using Monte Carlo simulation.

Keywords: classification; data mining; machine learning; Monte Carlo simulation; TOPSIS

INTRODUCTION

The technique for order preference by similarity to ideal solution (TOPSIS) is a classical method to solve multicriteria decision-making (MCDM) problems first developed by Hwang and Yoon (1981), subsequently discussed by many (Chu, 2002; Olson, 2004b; Peng, 2000). TOPSIS is based on the concept that alternatives should be selected that have the shortest distance from the positive ideal solution (PIS) and the farthest distance from the negative ideal solution (NIS), or nadir. The PIS has the best measures over all attributes, while the NIS has the worst measures over all attributes.

Multicriteria decision-making (MCDM) problem recently has received attention from artificial intelligence, machine learning, and data mining communities (Arie & Leon, in press; Spathis, Doumpos, & Zopounidis, 2002; Zopounidis & Doumpos, 1999). Based on
preference disaggregation approach estimates, a set of additive utility functions and utility profiles using linear programming techniques, Zopounidis and Doumpos (1999) present an application of the Utilities Additives Discriminantes (UTADIS) method in real-world classification problems concerning the field of financial distress. Spathis et al. (2002) proposed a multicriteria decision aid method for an innovative classification methodology in detecting firms’ falsified financial statements (FFS) and in identifying the factors associated with FFS. The proposed method is believed to outperform traditional statistical techniques on a sample of 76 Greek firms.

As an MCDM technique, TOPSIS also provides a mechanism that is attractive in data mining (Olson & Wu, 2005), because it can consider a number of attributes in a systematic way without very much subjective human input. Data, whether discrete or continuous, are standardized to a range between 0 and 1. TOPSIS does include weights over the attributes that are considered. However, such weights can be obtained through regression of standardized data where measurement scale differences are eliminated (Olson, 2004b). This allows machine learning in the sense that data can be analyzed without subjective human input. This article demonstrates the method to automatically classify credit score data into groups of high expected repayment and low expected repayment, based upon the concept of TOPSIS.

**TOPSIS FOR DATA MINING**

The overall approach is to begin with a set of data, which, in traditional data mining practice, is divided into training and test sets. Data may consist of continuous or binary numeric data, with the outcome variable being binary. A training data set is used to identify maximum and minimum measures for each attribute. The training set then is standardized over the range of 0 to 1, with 0 reflecting the worst measure and 1 the best measure over each attribute. Then, relative weight importance is obtained by regression over the standardized data in order to explain outcome performance in the training data set. (An intermediate third data set could be created for generation of weights, if desired.)

**TOPSIS DATA MINING METHOD**

The algorithm we propose consists of following steps:

**Step 1: Data Standardization**

In accordance with the prior presentation, training data set is standardized so that each observation \( j \) over each attribute \( i \) is between 0 and 1. Let the decision matrix \( X \) consist of \( m \) indicators over \( n \) observations. The normalized matrix transforms the \( X \) matrix. For indicator \( i = 1 \) to \( m \), identify the minimum \( x_{ij}^- \) and the maximum \( x_{ij}^+ \). Then, each observation \( x_i^j \) for \( j = 1 \) to \( n \) can be normalized by the following formulas:

For measures to be maximized:

\[
y_i^j = \frac{x_i^j - x_i^-}{x_i^+ - x_i^-}
\]

(1)

For measures to be minimized:

\[
y_i^j = 1 - \frac{x_i^j - x_i^-}{x_i^+ - x_i^-}
\]

(2)

which yields values between 0 (the worst) and 1 (the best).

**Step 2: Determine Ideal and Nadir Solutions**

The ideal solution consists of standardized values of 1 over all attributes, while the nadir solution consists of values of 0 over all attributes.

**Step 3: Calculate Weights**

In decision analysis, these weights would reflect relative criterion importance (as long as scale differences are eliminated through standardization). Here, we are interested in the relative value of each attribute in explaining the outcome of each case. These \( m \) weights \( w_i \)...
A Comparative Study of Associative Classifiers in Mesenchymal Stem Cell Differentiation Analysis
[www.igi-global.com/chapter/comparative-study-associative-classifiers-mesenchymal/73479?camid=4v1a](www.igi-global.com/chapter/comparative-study-associative-classifiers-mesenchymal/73479?camid=4v1a)

Visualization in Learning: Perception, Aesthetics, and Pragmatism
[www.igi-global.com/chapter/visualization-in-learning/150180?camid=4v1a](www.igi-global.com/chapter/visualization-in-learning/150180?camid=4v1a)