International Journal of Systems Science
Vol. 37, No. 13, 20 October 2006, 981–986

Taylor & Francis
Taylor & Francis Group

Data mining and simulation: a grey relationship demonstration

DESHENG WU†¶*, DAVID L. OLSON‡ and ZHAO YANG DONG§

†Rotman School of Management, University of Toronto, Toronto, Ontario, Canada M55-3E6
‡Department of Management, University of Nebraska, Lincoln, NE 68588-0491, USA
§School of Information Technology and Electrical Engineering, University of Queensland,
St. Lucia, QLD 4072, Australia
¶School of Business, University of Science and Technology of China, He Fei 230026,
An Hui Province, P.R. China

(*Received 15 July 2005*)

Fuzzy data has grown to be an important factor in data mining. Whenever uncertainty exists, simulation can be used as a model. Simulation is very flexible, although it can involve significant levels of computation. This article discusses fuzzy decision-making using the grey related analysis method. Fuzzy models are expected to better reflect decision-making uncertainty, at some cost in accuracy relative to crisp models. Monte Carlo simulation is used to incorporate experimental levels of uncertainty into the data and to measure the impact of fuzzy decision tree models using categorical data. Results are compared with decision tree models based on crisp continuous data.

Keywords: Data mining; Uncertainty; Fuzzy sets; Monte Carlo Simulation

1. Introduction

Data mining is an emerging area of computational intelligence. Modern organizations face a serious challenge in how they should make decisions from massively increased information so that they can better understand their markets, customers, suppliers, operations, and internal business processes. The field of data mining aims to improve decision-making by focusing on discovering valid, comprehensible, and potentially useful knowledge from large data sets.

This article presents a demonstration of the use of Monte Carlo simulation in grey related analysis for data mining purpose. Simulation is used to incorporate varying levels of uncertainty into a representative data set for customer segmentation. We used common random number seeds in generating data. Simulation also provides a means to more completely describe expected results, to include identification of the probability of a particular option being best in a multi-attribute setting. Use of simulation to evaluate model results in fuzzy data mining applications include those by Hu *et al.* (2004), Lee and Lee (2004), Kuo *et al.* (2005a, 2005b), Hu (2005), and Chen and Huang (2005). The Monte Carlo simulation is useful in designing experiments to verify the proposed data mining algorithms in these literatures. Kuo *et al.* (2005a, 2005b) employ Monte Carlo simulation to generate various artificial data sets and then with these artificial data sets verify the two proposed methods: The first is ant K-means algorithm by modifying the K-means as locating the objects in a cluster with the probability; the second is an integration of self-organizing feature mapping and genetic K-means. Krolzig and Hendry (2001) designed several Monte Carlo simulation experiments to select econometric models from a computer automation perspective, focusing on general-to-specific reductions. Hu (2005) design a fuzzy classifier as a fuzzy information retrieval system to mine ‘‘useful’’ or ‘‘meaningful’’ fuzzy concepts from training patterns for a classification problem. This fuzzy classifier performs quite well in his designed simulation experiment.

In this article, we will utilize Monte Carlo simulation for data mining purposes in a way different from those in the aforementioned literatures. We apply

*Corresponding author. Email: dwu@rotman.utoronto.ca

International Journal of Systems Science
ISSN 0020–7721 print/ISSN 1464–5319 online © 2006 Taylor & Francis
http://www.tandf.co.uk/journals
DOI: 10.1080/00207720600891521

Monte Carlo simulation to results of decision tree analysis of real credit card data. Both crisp (continuous data) and fuzzy (categorical data) decision tree models are applied to the same data set. Relative performances of crisp and fuzzy decision tree models are assessed in the conclusions. The following section discusses means to model uncertainty in data in order to include Monte Carlo incorporation of uncertainty. Section 3 discusses the decision tree models for both crisp and fuzzy data. The data set itself, simulated customer expenditure data, is described along with fuzzy categories and the experimental procedure. Section 4 analyzes results, and Section 5 presents the conclusions.

2. Modeling

Decision-making with uncertainty has progressed in a variety of directions throughout the world, leading to the development of probability theory (Pearl, 1988), fuzzy theory (Dubois and Prade, 1980), rough sets (Pawlak, 1982), grey sets (Deng, 1982), and vague sets (Gau and Buehrer, 1993). Real-life decisions usually involve high levels of uncertainty. Decision-making methods with uncertain input in the form of fuzzy sets and rough sets have been widely published in data mining (Witold, 1998; Hu *et al.*, 2003). The method of grey analysis (Deng, 1982) is an approach reflecting uncertainty. This article discusses the use of a common data analysis technique, i.e., Monte Carlo simulation, to this model to reflect uncertainty as expressed by fuzzy inputs. Monte Carlo simulation has been used to evaluate data mining (for instance, Rocco and Claudio, 2003), but the application here focuses on deeper analysis of decision tree rules in fuzzy (grey related) domains. While the example used is on an artificial data set, this data set represents a real application of great importance. Monte Carlo simulation provides a way to display the output of a data mining model and a means to reflect relative accuracy of alternative models under conditions of error in measures.

2.1 *Grey related analysis*

Grey related analysis is a decision-making technique that can be used to deal with uncertainty in forms of fuzzy data. Suppose that a multiple-attribute decision-making problem with interval numbers has m feasible plans $X_1, X_2, \ldots, X_m$, n indexes, and uncertain weight value w_j of index G_j, but $w_j \in [c_j, d_j]$, $0 \leq c_j \leq d_j \leq 1$, $j = 1, 2, \ldots, n$, $w_1 + w_2 + \cdots w_n = 1$, and the index value of jth index G_j of feasible plan X_i is an interval number $[a_{ij}^-, a_{ij}^+]$, $i = 1, 2, \ldots, m$, $j = 1, 2, \ldots, n$. When $c_j = d_j$, $j = 1, 2, \ldots, n$, the multiple attribute decision-making problem with interval numbers is called a multiple-attribute decision-making problem with interval-valued indexes. When $a_{ij}^- = a_{ij}^+$, $i = 1, 2, \ldots, m$, $j = 1, 2, \ldots, n$, the multiple-attribute decision-making problem with interval numbers is called a multiple-attribute decision-making problem with interval-valued weights. Using grey related analysis for decision-making with fuzzy data generally follows these steps:

Step 1. Construct decision matrix A with index number of interval numbers.

Step 2. Transform "contrary index" into positive index.

Step 3. Standardize decision matrix A with index number of interval numbers, obtaining standardizing decision matrix.

Step 4. Calculate interval-number-weighted matrix.

Step 5. Determine reference number sequence, which is composed of the optimal weighted interval number index value for every alternative.

Step 6. Calculate connections between alternatives using the reference number sequence.

Step 7. Determine optimal alternative with largest value of connection. Now, the principle and steps of the grey related analysis method are demonstrated by the following case illustration. For another route of demonstrating grey related analysis and simulation, the readers might refer to our recent work (Olson and Wu, 2006).

2.2 *Monte Carlo simulation*

Monte Carlo simulation provides an analytical method meant to imitate a real-life system in case other analyses are too mathematically complex or too difficult to reproduce. To deal with both fuzzy weights and fuzzy alternative performance scores over attributes, we develop a Monte Carlo simulation model of this decision. We implement our Monte Carlo simulation in the spreadsheet, which randomly generates values for uncertain variables over and over to simulate a grey related decision tree model described in the following. The simulation was controlled, using 10 unique seed values to ensure that the difference in simulation output due to random variation was the same for each alternative.

3. Grey related decision tree model

Grey related analysis is expected to provide improvement over crisp models by better reflecting the uncertainty inherent in many analysts' minds.

Data mining models based on such data are expected to be less accurate, but hopefully not by very much. However, grey related model input would be expected to be stabler under conditions of uncertainty where the degree of change in input data increased.

3.1 *Decision tree models*

We applied decision tree analysis to a set of consumer expenditure data (10,000 observations total), reflecting a small portion of the type of data that is widely used for customer segmentation analysis. There is one output variable (the proportion of income spent on groceries) that was converted to a binary variable, with 1 representing greater than average expenditure and 0 representing below average expenditure on groceries. This data was converted to grey by grouping it into ranges representing cut-offs between fuzzy categories. There were 13 available explanatory variables (table 1), which include the cutoff limits between fuzzy categories.

The variable churn was the number of credit card balances canceled over the last 12 months. The data set was divided in two, using the first 5000 observations for building decision tree models and the last 5000 for testing. The nature of the dividing point for the proportion of income spent on groceries led to a balanced data set. The categories given in table 2 give the cutoffs for fuzzifying the continuous numbers into categorical data.

PolyAnalyst software was used to build the decision trees. Since the data was controlled (using the same 5000 observations for training each model), different minimum support levels were used to generate models. PolyAnalyst also allowed a setting of optimistic (tending to yield more rules) and pessimistic (tending to reduce the number of rules). The pessimistic setting was used, yielding a total of three combinations of settings that were used for both continuous and categorical data. Table 2 gives the model yielded for each of the minimum support settings for the categorical data.

Table 2 demonstrates that the categorical model had a grand total of 9 leaves, and a depth of 3 (it used only three explanatory variables). The continuous models generated more rules. Table 3 reports these results.

3.2 *Simulation analysis of output*

These models were then entered into an Excel spreadsheet containing the 5000 test observations. For each of the variables used in the decision, Monte Carlo simulation (supported by Crystal Ball software) was applied. A perturbation of each input variable was generated, set at five different levels of perturbation by varying the level of the standard deviation. The intent was to measure the loss of accuracy for crisp (continuous) and grey related (categorical) models. Age, dependents, and income were continuous (but integer), and dispersed values were simulated using the integer of a normal distribution with a mean 0 and the standard deviation given in table 4. Marital status consisted of coded values, and the proportion of membership in each

Table 2. Categorical decision tree.

Dependents	Income	Marital status	Grocery spending
None	High		Low
None	Mid	Divorced	High
None	Mid	Married or single	Low
None	Low	Divorced or married	High
None	Low	Single	Low
One	High		Low
One	Mid or low	Divorced or married	High
One	Mid or low	Single	Low
Multiple			High

Table 1. Available explanatory variables.

Variable	Data type	Category 1	Category 2	Category 3
Age	Integer	< 32	32 to 48	49 and up
Gender	Binary	0 – female	1 – male	
Marital status	Categorical	Single	Divorced	Married
Dependents	Integer	0	1	Multiple
Income	Continuous	<34,000	34,000 to 55,000	>55,000
Years on job	Integer	<5	5 to 10	>10
Years in town	Integer	<1	1 to 5	>5
Years education	Integer	<12	12 to 15	16 or more
Drivers license	Binary	0 – no	1 – yes	
Own home	Binary	0 – no	1 – yes	
Number of credit cards	Integer	0	1	Multiple
Churn	Integer	0	1	Multiple

Table 3. Model metrics.

	Categorical	Continuous MS10	Continuous MS100	Continuous MS200
Leaves (rules)	9	14	11	14
Depth	3	6	7	6
Variables	Dependents Income Marital status	Dependents Income Marital status Age	Dependents Income Marital status Age	Dependents Income Marital status Age

Table 4. Simulated parameters to reflect dispersion.

Standard deviations	Age	Marital status	Dependents	Income
0	0	0	0	0
0.25	4.25	0.025	0.25	5.5
0.5	8.5	0.05	0.5	11.0
1.0	17.0	0.10	1.0	22.0
2.0	34.0	0.20	2.0	44.0
3.0	51.0	0.30	3.0	66.0

was varied if a uniform number was drawn less than the given parameter. This scheme allowed controlling the random variation so that the continuous and categorical models reflected the same relative changes. Continuous test data were randomly varied, and then the implied categorical assignments made.

4. Results

The test data were then simulated using Crystal Ball software, an add-in to Excel. Each model was simulated 1000 times using common random number seeds. The outcome measurement was the proportion of correct classifications. The results were fairly evenly balanced between type I and type II errors, so combined model accuracy over the test set was used. Figure 1 shows the histogram of correct proportion for the categorical model.

Table 5 gives the minimum and maximum proportion correctly predicted for each of these models.

Figure 2 compares these numbers for the categorical model and the MS10 continuous model.

Figure 2 shows that, with no variance in the data, the continuous regression trees were more accurate. But it also shows that the categorical model became as accurate as the continuous model at very high levels of data uncertainty (SD = 3). This behavior was typical of all three levels of minimum support used. However, faced with uncertain data, the categorical model is less affected.

Monte Carlo simulation was useful in more completely describing the variation in accuracy due to the uncertainty in input data. It could also be used to identify the probability of a specific observation being assigned to one of two or more output categories. Past applications studied (Olson and Wu, 2005) involved small models and little computation time. In the application in this study, the data set was a bit larger and the computational time was quite extensive. However, simulation is the most flexible of analytic techniques, and our intent is to demonstrate the potential value of Monte Carlo simulation in analyzing data mining output.

5. Conclusions

Data mining involves many challenging tasks including handling of qualitative attributes, exploitation of large data sets for model development goal by efficient computation procedures, and derivation of easily understandable decision models. Fuzzy set theory is especially useful to deal with problems with qualitative attributes. Monte Carlo simulation provides value in giving a more complete picture of the dispersion of results obtained and provides probabilistic explanations that are useful for decision making.

This study has demonstrated an approach using Monte Carlo simulation and grey related analysis by taking into account fuzzy data. Monte Carlo Simulation is used as a data mining technique to measure the impact of fuzzy decision tree models (using categorical data) compared to decision tree models based on continuous data. The proposed approach is applied to a case with a typical business data set used for data mining.

The overall conclusion is that, if the data included no uncertainty, fuzzification as applied here reduces accuracy slightly. But as higher levels of uncertainty are

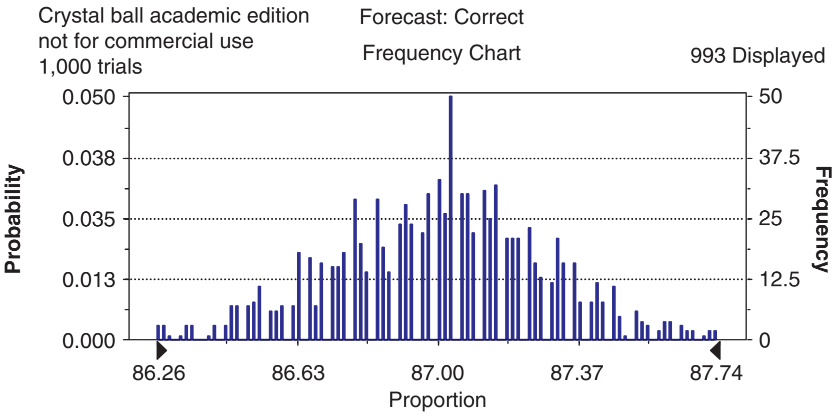

Figure 1. Histogram of categorical model correct predictions

Table 5. Correct prediction ranges.

Standard deviations	Categorical All 3 MS levels	Continuous MS10	Continuous MS100	Continuous MS200
0	0.8552	0.8982	0.8982	0.8954
0.25	0.8332–0.8488	0.8610–0.8810	0.8590–0.8804	0.8326–0.8546
0.50	0.8062–0.8310	0.8264–0.8504	0.8246–0.8524	0.8266–0.8490
1	0.7630–0.7964	0.7738–0.8084	0.7334–0.8016	0.7706–0.8040
2	0.6916–0.7298	0.7034–0.7388	0.7014–0.7410	0.7040–0.7380
3	0.6618–0.6986	0.6656–0.7014	0.6668–0.7050	0.6682–0.7038

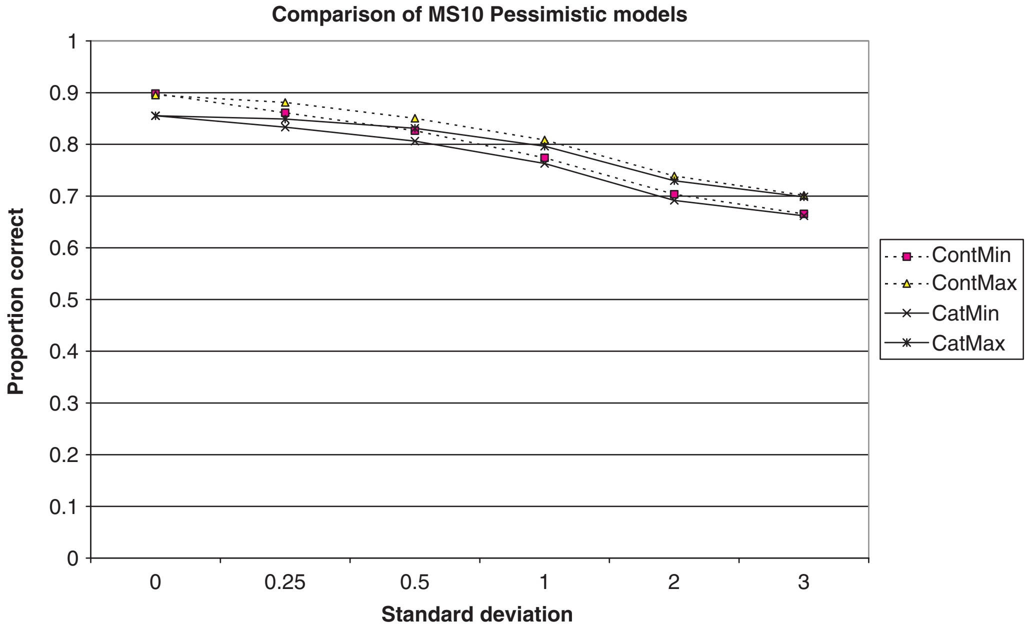

Figure 2. Comparison of MS10Pess models

present in the data, fuzzy data approaches the accuracy of crisp data. Both would degrade in accuracy as uncertainty was increased, as would be expected.

References

Y.-L. Chen and T.C.-K. Huang, "Discovering fuzzy time-interval sequential patterns in sequence databases", *IEEE Trans. Sys. Man & Cybernetics Part B*, 35(5), pp. 959–972, 2005.

J.L. Deng, "Control problems of grey systems", *Syst. Control. Lett.*, 5, pp. 288–294, 1982.

D. Dubois and H. Prade, *Fuzzy Sets and Systems: Theory and Applications*, New York: Academic Press, Inc., 1980.

W.L. Gau and D.J. Buehrer, "Vague sets", *IEEE Trans, Sys. Man & Cybernetics*, 23, pp. 610–614, 1993.

Y.-C. Hu, "Finding useful fuzzy concepts for pattern classification using genetic algorithm", *Information Sciences*, 175(1/2), pp. 1–19, 2005.

Y. Hu, R. Chen and G. Tzeng, "Finding fuzzy classification rules using data mining techniques", *Pattern Recogn. Lett.*, 24(1–3), pp. 509–519, 2003.

Y.-C. Hu, J.-S. Hu, R.-S. Chen and G.-H. Tzeng, "Assessing weights of product attributes from fuzzy knowledge in a dynamic environment", *Eur. J. Oper. Res.*, 154(1), pp. 125–143, 2004.

W.-J. Lee and S.-J. Lee, "Discovery of fuzzy temporal association rules", *IEEE Trans. Sys., Man & Cybernetics, Part B*, 34(6), pp. 2330–2342, 2004.

D.L. Olson and D. Wu, "Decision making with uncertainty and data mining". *Advanced Data Mining and Applications: First International Conference, ADMA 2005*, X. Li, S. Wang, Z.Y. Dong eds., Lecture Notes in Artificial Intelligence. Berlin: Springer 2005, pp. 1–9.

D.L. Olson and D. Wu, "Simulation of fuzzy grey relationships". *Eur. J. Oper. Res.*. 2006 (in press).

Z. Pawlak, "Rough sets", *Int. J. Inf. & Comput. Sci.*, 11, pp. 341–356, 1982.

J. Pearl, "Probabilistic reasoning in intelligent systems, networks of plausible inference", Morgan Kaufmann, San Mateo, CA 1988.

S. Rocco and M. Claudio, "A rule induction approach to improve Monte Carlo system reliability assessment", *Reliab. Eng. Syst. Safe.*, 82(1), pp. 85–92, 2003.

P. Witold, "Fuzzy set technology in knowledge discovery", *Fuzzy Sets and Systems*, 98(3), pp. 279–290, 1998.

R.J. Kuo, H.S. Wang, T. Hu and S.H. Chou, "Application of ant K-means on clustering analysis", *Comput. & Math. Appl.*, 50(10–12), pp. 1709–1724, 2005a.

R.J. Kuo, J.L. Liao and C. Tu, "Integration of ART2 neural network and genetic K-means algorithm for analyzing Web browsing paths in electronic commerce", *Decision Support Systems*, 40(2), pp. 355–374, 2005b.

H. Krolzig and D.F. Hendry, "Computer automation of general-to-specific model selection procedures", *J. Econ. Dyn. Control*, 25(6–7), pp. 831–866, 2001.