INFORMATICA, 2005, Vol. 16, No. 4, 557–570

Maximum Resolution Topology for Multi-Attribute Dichotomies

James K. HO
Information & Decision Sciences, College of Business Administration
University of Illinois at Chicago
Chicago, IL 60091 USA
e-mail: jimho@uic.edu

Sydney C.K. CHU
Operations Research Group, Department of Mathematics, University of Hong Kong
Hong Kong SAR, China
e-mail: schu@hku.hk

Received: February 2005

Abstract. The Star Plot approach to high-dimensional data visualization is applied to multi-attribute dichotomies. It is observed that the areas of the plot for the two parts of a dichotomy may be used as an aggregate measure of their relative dominance. An optimization model is developed to determine a topology (or weighted configuration of the attributes) that maximizes the resolution of this measure with respect to a given set of reference dichotomies.

Key words: multi-attribute dichotomy, data visualization, aggregate measure, value function, optimization, exploratory data analysis, data mining.

1. Introduction

A multi-attribute dichotomy is any multi-dimensional dataset in which the dimensions can be partitioned into two groups, each considered to be contributing to one part of the dichotomy. For example, the set of options in a decision analysis may have cost and benefit factors. In the clinical study of certain medical conditions, there may be a genetic-environmental dichotomy (McGue, 1997). For consumer attitudes in market research, a key issue can be whether customers are more price or quality focused (Riesz, 1978). With online auction markets, one can study whether conditions are favorable to either buyers or sellers (Ho, 2004). In each case, it is of interest to classify instances of the dataset according to the dominance of one part or the other of the dichotomy. In decision analysis, this identifies cost-effective options. In medical diagnosis, such classification may help to select treatment methods. In marketing campaigns, proper segmentation of customers can guide the balance between pricing or product improvement incentives (Johnson, 2000).

Many methods have been explored and well studied for the problem of aggregating multiple and incommensurate attribute functions. The aggregate measure may assume

simple additive forms for which appropriate weights must be determined (Poyhonen and Hamalainen, 2001) through some process of subjective preference evaluation by some decision maker (Dyer and Sarin, 1979; Kirkwood and Sarin, 1980). In the more comprehensive theory of multi-attribute utility, the decision maker's attitude towards risk is taken into consideration by the evaluation of uncertain options among outcomes (Keeney and Raiffa, 1976). The approach developed in this work is a departure from such classical methodology. While an aggregate measure is sought to discern the relative dominance of the two parts of a dichotomy, no pair-wise comparison of outcomes, or preference evaluation under uncertainty is invoked. Instead, a subset of the dichotomies with "prejudged" dominance of one part or the other is used as a frame of reference. An aggregate measure is then derived from an optimization model which provides maximum resolution in discerning dominance with respect to the reference set.

In this regard, our approach is grounded in Exploratory Data Analysis (EDA) (Tukey, 1977). In contrast to traditional inferential statistics, which focuses on hypothesis testing designed to verify a priori hypotheses about relations among variables, EDA does not presume any model on such relationships. It is a philosophy, rather than a collection of techniques, to use primarily graphical visualization to gain insight into a data set. With open-minded exploration, the goal is to uncover hidden structure and systematic patterns that may lead to further discovery of knowledge. While EDA was pioneered in the days of hand-drawn graphics, advances in database technology and computer graphics have enabled more sophisticated specializations. EDA on large and complex databases now falls under the rubric of Data Mining (Chen *et al.*, 1996), and computerized graphical methods are known as Data Visualization (Hoffman and Grinstein, 2001; Keim, 2002).

The major objective of Data Mining is Knowledge Discovery in Databases, one popular definition of which is the non-trivial process of identifying valid, novel, potentially useful, and understandable patterns in data. One major topic is the discovery of association rules, which shed light on correlations in the data. Such knowledge has practical use for decision support when combined with appropriate decision-analytic models (Bohanec and Zupan, 2001). Another major topic is classification, which subsumes the more traditional subjects of cluster analysis and pattern recognition (Ankerst *et al.*, 2000). In this work, the Star Plot Method (Chambers *et al.*, 1983) of high-dimensional data visualization is applied to multi-attribute dichotomies. Beyond giving shape to the dichotomies, hence enabling visual EDA, the respective areas of the two parts suggest a plausible aggregate measure for classifying the dichotomies according to the dominance of their parts. A mathematical programming model is formulated to produce an optimal topology for visualizing the dichotomies, so that the areas in the plot provide an aggregate measure with maximum resolution with respect to some reference set. In this context, the proposed approach is a novel application of optimization in visual data mining.

Section 2 motivates topological analysis of multi-attribute dichotomies, and describes the application of the Star Plot Method. The optimization model for a maximum resolution topology is introduced in Section 3. An illustrative example in classifying the cost-effectiveness of Law Schools is given in Section 4. Preliminary results of applications and directions for future work are discussed in Sections 5 and 6.

2. Topological Model for Visualization of Multi-Attribute Dichotomies

2.1. *Star Plot for High-Dimensional Data*

One of the simplest yet visually most powerful display of multi-dimensional data is known as the star plot (Chambers *et al.*, 1983). The values for an instance in the dataset are mapped on symmetrically placed axes representing the dimensions radiating from a common origin. The polygonal glyph obtained by connecting the plotted values gives "shape" to the data. Pattern and structure of such shapes can then provide useful visual cues for EDA. Chambers *et al.* (1983) presented what has become the best known example, which was based on automobile data in 1979. There are twelve dimensions and 74 instances in the dataset. The dimensions are:

1. PRICE
2. MILEAGE (MPG)
3. REPAIR RECORD (1978) 1 = WORST & 5 = BEST
4. REPAIR RECORD (1977) 1 = WORST & 5 = BEST
5. HEAD ROOM (INCHES)
6. READ SEAT ROMM (INCHES)
7. TRUNK SPACE (CUBIC FEET)
8. WEIGHT (POUNDS)
9. LENGTH (INCHES)
10. TURNING CIRCLE (FEET)
11. DISPLACEMENT (CUBIC INCHES)
12. GEAR RATIO

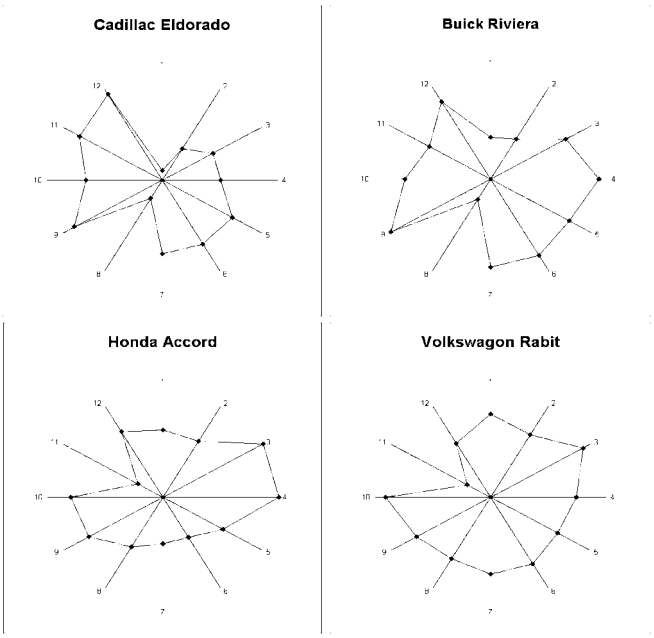

Fig. 1. Star plots of 1979 automobile data.

With appropriate scaling, all the attributes can be plotted with increasing value being desirable. The star glyphs for four of cars in the dataset: Cadillac Eldorado, Buick Riviera, Honda Accord, and Volkswagon Rabit are shown in Fig. 1.

Even from this small sample, it can be observed that larger American models exhibit different glyph shapes from compact imports. More recent and significant applications include data mining of molecular databases in automated drug discovery research (Song *et al.*, 2002), and international comparison of service use profiles of mental health patients (Patel, 2003).

2.2. *Star Plot for Multi-Attribute Dichotomies*

While it is obvious and well known that the shape of the star glyphs depends on the configuration of the attributes along the radial dimensions of the plots, EDA to date simply encourages analysts to investigate the variations on an ad hoc basis. A first step toward a more systematic approach is motivated by the topological analysis of online auction markets (Ho, 2004). To address the question of "What shape is an online auction market in ?" twelve attributes are identified:

1. NET ACTIVITY (auctions with bids)
2. PARTICIPATION (average number of bids per auction)
3. SELLER DIVERSITY (distribution of offers)
4. SELLER EXPERIENCE (distribution of sellers' ratings)
5. MATCHING (auctions ending with a single bid)
6. SNIPING (last minute winning bids)
7. RETAILING (auctions ending with the Buy-It-Now option)
8. BUYER DIVERSITY (distribution of bidder participation)
9. BUYER EXPERIENCE (distribution of buyers' ratings)
10. DUELING (evidence of competitive bidding)
11. STASHING (evidence of stock-piling)
12. PROXY (use of proxy bidding as evidence of true valuation)

Attributes 1 and 2 are common parameters of the market and considered neutral to buyers and sellers. Attributes 3 through 7 reflect market conditions for buyers in the sense that higher observed values are deemed to be advantageous. Similarly, attributes 8 through 12 are considered to be seller dimensions, as for example, higher level of competitive bidding in the form of head-to-head dueling is a welcome condition for sellers. By arranging the seller attributes on the left side, and buyer attributes on the right side of a star plot, with the neutral attributes on the top and bottom verticals, respectively, a glyph for the buyer-seller dichotomy can be drawn. Mining data on completed auctions available on eBay.com, the largest platform for online auctions over the Internet, four example markets in June and July 2004 are shown in Fig. 2. These markets: digital cameras (medium capacity, 3–4 Mega pixels), diamond rings, classical music CDs, and toys and games with the Star Wars theme, represent a small but diverse sample of the over 25,000 categories being offered on eBay. The star plots, which are snapshots of the markets at a particular time point, illustrate that different markets take on different shapes according to their operational characteristics at the moment.

In general, for dichotomies with no convenient "neutral" attributes to serve as the separating vertical axis, two dummy dimensions are used. These may be labeled, say,

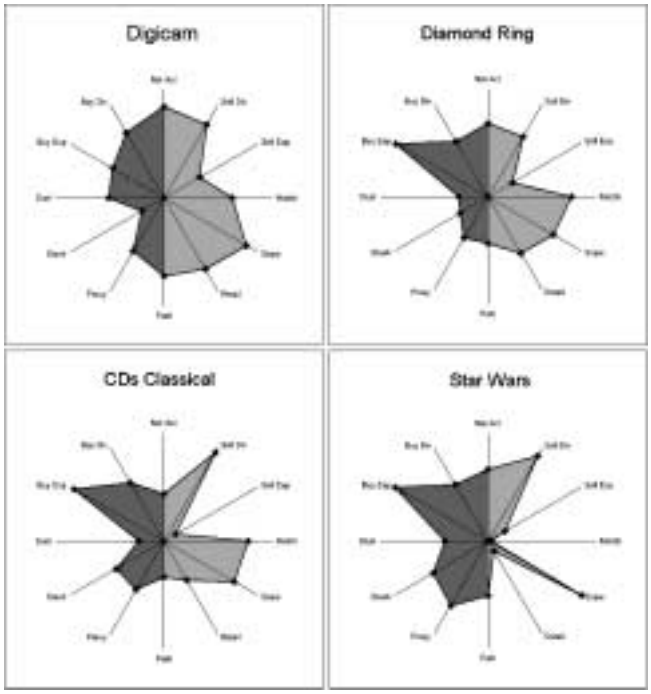

Fig. 2. Star plots for four online auction markets.

"Top" and "Bottom" and set at arbitrary constant values, or at the mean of all attribute values for the data instance.

2.3. *Topological Model for Multi-Attribute Dichotomies*

Given the star glyph of a multi-attribute dichotomy, as exemplified in Fig. 2, it will be both visually and intuitively appealing if the areas covered by the two parts can be used as a meaningful aggregate measure of their relative dominance. A larger area on the left side of the glyph means dominance by the left part, and vice versa. In the case of online auction markets, this can be interpreted as market conditions being advantageous to either buyers or sellers. In mathematical terms, the aggregate value function takes the form of the sum of pairwise products of adjacent attributes:

$$V(X_1, \ldots, X_n) = C \sum X_i X_j,$$

where attributes i and j are adjacent; X_i is the value of attribute i, $i = 1, \ldots, n$; and C is some scaling constant.

While plausible, since increasing value of an attribute contributes positively to its designated part, as well as the latter's area in the glyph, the concept can be refined to realize its potential.

To this end, the first step is to render the glyph unit free by normalizing the data on each dimension to the unit interval [0, 1]. The second step is to render the glyph context free by harmonizing the dimensions as follows. For each attribute, the quartiles for the values in the entire dataset are computed. A spline function (Cline, 1974) is constructed to map these quartiles into the [0.25, 0.5, 0.75] points of the unit interval. This way, a hypothetical data instance with all attributes at mean values of the dataset will assume the shape of a symmetrical polygon with vertices at the mid-point of each radial axis. In this frame of reference, all shapes and sizes are relative to this generic "average" glyph, and free of either units or specific context of the attributes. For our exploratory work, simple second-order (piecewise linear) splines are used.

Finally, the critical potential in the concept of using area as aggregate measure arises from the degrees of freedom allowed by the topology of the glyph, namely, the configuration of the attributes, and the angles between adjacent pairs thereof. For any given arrangement of the attributes, the standard star plot produces a glyph along symmetrically spaced radial axes. Variations from this symmetry imply a feasible set of shapes and areas, which along with permutations of the configuration, offer the choice of topologies that may suit further criteria for a meaningful aggregate measure function. In particular, we use a subset of the data instances, for which the dichotic dominance has been prejudged, in an optimization model to derive a topology with maximum resolution in discerning dominance with respect to the reference subset. The aggregate value function now takes the form of the sum of weighted pairwise products of adjacent attributes

$$V(X_1, \ldots, X_n) = C \sum w_{ij} X_i X_j,$$

where attributes i and j are adjacent, and w_{ij} is the weighing factor that determines the angle between them.

Several options are open in choosing the reference subset of dichotomies. In the absence of significant insight into the classification of the dichotomies, and especially in the early stages of EDA, any initial dataset can be used on an ad hoc basis. An arbitrary configuration of the attributes within each part of the dichotomy is selected and evenly spaced. The classification of the dichotomies according to this topology is fed into the optimization model. The resulting optimal topology with respect to this reference set provides a working definition of dominance for the dichotomies. This is analogous to selecting a portfolio of stocks to provide an index for a stock market. The performance of any stock can be gauged relative to the index, which may be arbitrarily chosen initially. With better knowledge of the significance of individual stocks, more useful indices can be established. By the same token, the maximum resolution topology for multi-attribute dichotomies can be adaptively refined as the EDA progresses.

When there is expert opinion, or a role for a decision maker, the reference subset can be chosen subjectively. Note that rather than eliciting preference evaluation as in other approaches to aggregate multi-attribute value functions, only judgment on the classification of a subset of data instances is required. In practice, the decision maker is asked to identify some dichotomies that he or she judges with confidence to be "left-dominant", and

others (preferably more or less equal in number) that are "right-dominant". Obviously, a certain degree of consistency is expected of the subjective judgment for the optimization to be feasible. Minor systematic discrepancies may be compensated by a uniform rescaling of the "left" and "right" attributes appropriately.

Yet another option to select the reference dichotomies can be an integral part of the EDA process. For example, for market segmentation in consumer attitude research, the dichotomy is cost focus versus quality focus. Based on responses to surveys that supply data on how customers rate various cost and quality attributes, the objective is to classify all potential customers. In this case, focus groups can be employed to identify customers who are bona fide specimens of each class. Their corresponding data points are then used as the reference subset in the optimization model.

Subject to the constraints of preserving the prejudged dominance in the reference subset of dichotomies, an optimal topology (configuration of attributes and angles between adjacent pairs) is sought that maximizes the discriminating power (sum of absolute differences in left and right areas) for the reference subset. Before presenting the details of the model, it should be remarked that maximum resolution in this context does not imply the classification of the most dichotomies. Any topology can classify all instances in a dataset (except in degenerate cases where the areas are numerically equal). The critical factor is that as aggregate measures, different topologies may classify the same dichotomy differently. The maximum resolution model rationalizes the choice for one that best "enforces" the classifications in the reference subset.

3. An Optimization Model for Maximum Resolution Topology

With the star glyph as the topology, we now introduce its optimization model. A (kth) data *record* consists of ("left" or) X-attribute data $\{x_i^k\}$ and ("right" or) Y-attribute data $\{y_j^k\}$, such that x_i^k denotes the kth record's ith X-attribute value and y_j^k denotes the kth record's jth Y-attribute value, with (unit interval data range)

$$\left\{0 \leqslant x_i^k \leqslant 1 | i \in I\right\} \quad \text{and} \quad \left\{0 \leqslant y_j^k \leqslant 1 | j \in J\right\},$$

where

$$I \equiv \left\{i = 1, 2, \cdots, m | \bar{x}_1 \geqslant \bar{x}_2 \geqslant \cdots \geqslant \bar{x}_{m-1}\right\}$$

and

$$J \equiv \left\{j = 1, 2, \cdots, n | \bar{y}_1 \geqslant \bar{y}_2 \geqslant \cdots \geqslant \bar{y}_{n-1}\right\}.$$

The $\bar{x}_i$ and $\bar{y}_j$ above are the average attribute-values over the records. The index sets I and J are defined as one convention to standardize a "default" configuration of the attributes. We assume the natural boundary conditions of

$$x_0^k = y_0^k \ (\geqslant) \ x_m^k = y_n^k \quad \forall k \in K$$

are assumed. These are either neutral attributes, or dummy attributes "Top" and "Bottom" assigned to the vertical axes as discussed in Subsection 2.2. The records are indexed by $k \in K$ in the manner to be explained below. When referring to the ith X-attribute as a variable rather than its value for a particular data instance, we shall write x_i instead of x_i^k; and similarly y_j instead of y_j^k.

In the star plot illustrated in Fig. 3, denote the *angles* between attributes x_{i-1} and x_i by α_i, $i \in I$, and that between attributes y_{j-1} and y_j by β_j, $j \in J$, for all $|K|$ records. Let the *weights* be

$$a_i \equiv \operatorname{Sin}\alpha_i, \;\; i \in I; \quad \text{and} \quad b_j \equiv \operatorname{Sin}\beta_j, \;\; j \in J.$$

And we have $0 \leqslant a_i \leqslant 1$, $i \in I$ and $0 \leqslant b_j \leqslant 1$, $j \in J$. Hence for the kth record, the sector area between attributes x_{i-1} and x_i, and that between attributes y_{j-1} and y_j, are given respectively by

$$A_i^k (= A_i^k(x)) \equiv \frac{1}{2} x_{i-1}^k x_i^k \operatorname{Sin}\alpha_i = \frac{x_{i-1}x_i}{2} a_i, \;\; i \in I;$$

and

$$B_j^k (= B_j^k(y)) \equiv \frac{1}{2} y_{j-1}^k y_j^k \operatorname{Sin}\beta_j = \frac{y_{j-1}y_j}{2} b_j, \;\; j \in J.$$

Next, we partition the data records according to their left-right dominance under equal weighting in a standard star plot. Using $a_i = 1/|I|$, and $b_j = 1/|J|$ in the above formulas for $A_i^k, \forall i$ and $B_j^k, \forall j$ respectively, we group the $|K|$ records into two partitions such that $K \equiv K^+ \cup K^-$, with

$$K^+ = \left\{k \in K | A^k \geqslant B^k\right\} \;\text{ and }\; K^- = \left\{k \in K | B^k > A^k\right\},$$

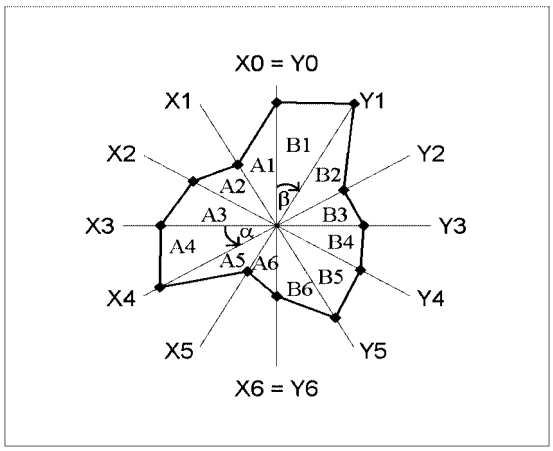

Fig. 3. Terminology for the optimization model.

where

$$A^k = \sum_{i \in I} A_i^k \text{ and } B^k = \sum_{j \in J} B_j^k.$$

Note that A^k is the area of the left part of the dichotomy in the star glyph spanned by its X-attributes, and B^k is the area of the right part of the dichotomy spanned by its Y-attributes of record k, $k \in K$. K^+ indicates records with X-attribute dominance (or left dominance), and K^- indicates records with Y-attribute dominance (or right dominance). With these notations, we formulate the *Maximum Resolution Topology* Linear Program (or MRT-LP for short) as follows.

$$\text{Max} \sum_{k \in K^+} [A_k - B_k] + \sum_{k \in K^-} [B_k - A_k].$$

Subject to $A^k - B^k \geqslant 0$, $k \in K^+$ and $B^k - A^k \geqslant 0$, $k \in K^-$

$$\sum_{i \in I} a_i = 1, \quad \sum_{j \in J} b_j = 1, \quad a_i \geqslant 0, \quad i \in I; \quad b_j \geqslant 0, \quad j \in J.$$

The objective function above, with each term constrained to be non-negative, measures the gross dominance by the larger part in the dichotomy. By maximizing, we seek the highest discriminating power, in an aggregate sense, to separate the parts.

Indeed, a numerical experiment with $|K| = 16$ records, $|K^+| = 8$, $|K^-| = 8$, $m = 3$, $n = 4$, and random input data for x_i and y_j over [0, 1], gives the following results. The maximum objective function value (=TTLMRT) is equal to 1.859, compared to the equal-weight total area difference given by

$$\sum_{k \in K^+} [A^k - B^k] + \sum_{k \in K^-} [B^k - A^k] \Big| \left\{ a_i = \frac{1}{|I|}, \forall i; b_j = \frac{1}{|J|}, \forall j \right\} = 1.278,$$

or an improvement of over 45%. However, the optimal weights (by a Lingo code) are: $a^* = (0.796, 0.000, 0.204)$ and $b^* = (0.000, 0.777, 0.223, 0.000)$. The extreme point solution by linear programming collapses three out of the seven inter-attribute angles, and cannot produce a proper star glyph displaying all the attributes. To circumvent this difficulty, we extend the MRT-LP model into its goal programming version (Chu, 2001) of MRT-GP by the addition of the usual deviation variables with total smoothing variation (=TTLVAR) and/or bounded deviation (=TTLDEV). This MRT-GP with decision variables a_i, $i \in I$ and b_j, $j \in J$ is then given by

Max	TTLMRT
Min	TTLDEV
Min	TTLVAR

Subject to $A^k - B^k = D^k$, $k \in K^+$ and $B^k - A^k = E^k$, $k \in K^-$

$$\sum_{i \in I} a_i = 1, \quad \sum_{j \in J} b_j = 1, \quad a_i \geqslant 0, \quad i \in I; \quad b_j \geqslant 0, \quad j \in J;$$

$$\text{TTLMRT} = \sum_{k\in K^+} D^k + \sum_{k\in K^-} E^k,$$
$$\text{TTLDEV} = \frac{1}{|I|}\sum_{i\in I}\big[\text{PDEVX}(i) + \text{NDEV}X(i)\big] + \frac{1}{|J|}\sum_{j\in J}\big[\text{PDEV}Y(j) + \text{NDEV}Y(j)\big],$$
$$\text{TTLVAR} = \frac{1}{|I|}\sum_{i\in I}\big[\text{PVAR}X(i) + \text{NVAR}X(i)\big] + \frac{1}{|J|}\sum_{j\in J}\big[\text{PVAR}Y(j) + \text{NVAR}Y(j)\big].$$

(Deviation bounds)

$$\frac{1}{|I|} - F * \frac{1}{|I|} \leqslant a_i \leqslant \frac{1}{|I|} + F * \Big(1 - \frac{1}{|I|}\Big), \quad i \in I,$$
$$\frac{1}{|J|} - F * \frac{1}{|J|} \leqslant b_i \leqslant \frac{1}{|J|} + F * \Big(1 - \frac{1}{|J|}\Big), \quad j \in J.$$

(Deviation goal constraints)

$$a_i - \text{PDEV}X(i) + \text{NDEV}X(i) = \frac{1}{|I|}, \quad i \in I,$$
$$b_j - \text{PDEV}Y(j) + \text{NDEV}Y(j) = \frac{1}{|J|}, \quad j \in J.$$

(Variation/smoothing goal constraints)

$$a_i - a_{i+1} - \text{PVAR}X(i) + \text{NVAR}X(i) = 0, \quad i \in I \qquad (a_{m+1} \equiv a_0),$$
$$b_j - b_{j+1} - \text{PVAR}Y(j) + \text{NVAR}Y(j) = 0, \quad j \in J \qquad (b_{n+1} \equiv b_0).$$

The $0 < F < 1$ parameter in the MRT-GP formulation above is the fraction of deviation from equal weights allowed for the weight variables a and b. An empirical trial, using weighted rather than pre-emptive GP with a deviation allowance of 50% ($F = 0.5$) returns 1.657 for TTLMRT, which is about a 30% improvement over 1.278. The weights are $a = (2/3, 1/6, 1/6)$ and $b = (1/8, 9/16, 3/16, 1/8)$.

This numerical instance illustrates only one of the two aspects of degrees of freedom of the topology – the "angles". The other aspect – "configuration" – can be simply treated by considering all the 144 ($= 3! \times 4!$) cases of complete permutations of $\{i \in I\}$ and $\{j \in J\}$ from the two sets of attributes. The example in Section 4 will demonstrate both in an actual application setting.

4. An Illustrative Example

To illustrate the optimization model for maximum resolution topology for multi-attribute dichotomies, we use an example in cost-benefit analysis of US law schools. The dataset, adapted from (ILRG, 1996), has 51 instances. Six attributes are used in a cost-benefit dichotomy. The cost attributes are Tuition, Room and Board, and Payment on Loan. The benefit attributes are ratings on Job Placement, Starting Salary, and Ranking (Reputation). The data is first normalized and harmonized as described in Subsection 2.3. The dummy dimensions "Top" and "Bottom" are set at the mean of all attribute values for the data instance. The subjective prejudgment by a decision maker to select the reference subset is then simulated. Under an arbitrary default topology (with the attributes ordered as listed above in the standard star plot), the five records that are most cost-dominant, and the five records that are most benefit-dominant are chosen, assuming that these are the "clear-cut" cases for the hypothetical decision maker. The data for this reference subset is shown in Table 1.

For each permutation of the configuration of the attributes, the MRT-GP problem comprises eight variables and 54 constraints. With six attributes (3 for each part), there are (3! x 3!) =36 permutations yielding four optimal configurations that are topologically equivalent due to symmetry. The average gain in resolution from optimization, as measured in Section 3, is 32.6% over all configurations. A maximum resolution topology for a typical data instance is shown in Fig. 4. The optimal configuration, counting from the "Top", is Room and Board, Tuition, and Loan Payment for the cost attributes; and Job Placement, Salary, and Ranking for the benefit attributes. The optimal angles (in units of π) are, clockwise from the "Top", (0.125, 0.625, 0.125, 0.125, 0.125, 0.625, 0.125, 0.125). Since the star glyph topology accentuates pairwise correlation among the attributes, we observe that Tuition and Loan Payment have the highest correlated effect for the cost attributes, whereas Placement and Salary do so for the benefit attributes, with each pair being assigned the largest weight of 0.625. For the data instance in Table 1, the aggregate measure for the cost attributes is 0.076, compared to that of 0.144 for the benefit

Table 1

The reference subset for cost-benefit dichotomy of law schools

Data ID	Tuition	R&B	Payment	Placement	Salary	Ranking
39	**0.612**	**0.356**	**0.454**	0.324	0.356	0.247
50	**0.785**	**0.815**	**0.464**	0.676	0.620	0.459
33	**0.643**	**0.642**	**0.409**	0.516	0.535	0.100
31	**0.690**	**0.404**	**0.585**	0.484	0.292	0.198
41	**0.617**	**0.501**	**0.711**	0.708	0.180	0/345
32	0.792	0.660	0.408	**0.740**	**0.660**	**0.786**
4	0.391	0.412	0.306	**0.580**	**0.388**	**0.622**
38	0.815	0.579	0.475	**0.868**	**0.548**	**0.753**
23	0.850	0.603	0.423	**0.676**	**0.660**	**0.802**
29	0.429	0.243	0.202	**0.548**	**0.532**	**0.231**

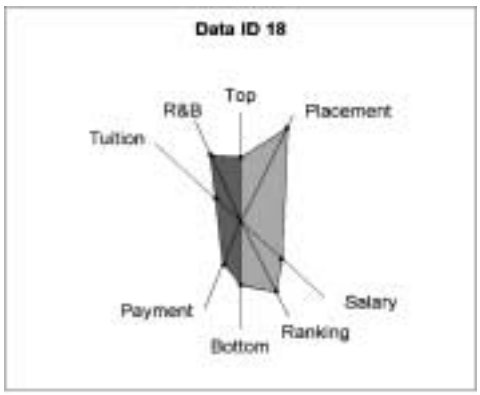

Fig. 4. Maximum resolution topology for law schools.

attributes, resulting in a classification of benefit-dominant. In other words, the law school represented by this data instance is deemed cost-effective using a maximum resolution aggregate measure induced by the decision maker's prejudgment on the reference cases. On this basis, of the entire data set of 51 law schools, 22 are rated cost-dominant, and 29 benefit-dominant.

5. Preliminary Results of Application

As examples of application of multi-attribute dichotomies, we summarize initial results in two ongoing research projects in economic analysis, and technological diffusion, respectively. Details of the studies will be issued separately in forthcoming reports.

The International Bank for Reconstruction and Development (The World Bank) has been gathering and analyzing data for insights into how investment climates vary around the world and how they influence growth and poverty. In particular, Investment Climate Surveys have been conducted in over 53 countries since 2001 to measure specific constraints facing firms, and relate them to measures of firm performance, growth, and investment (World Bank, 2004). The framework of Investment Climate Indicators (ICI) can be cast into a policy-resource dichotomy with *Policy Uncertainty*, *Licensing Regulations*, and *Tax Rate* as policy dimensions, and *Electricity*, *Cost of Financing*, and *Labor Supply* as resource dimensions. The goal in this application is to classify economies according to whether investment climate is constrained predominantly by either governmental policies, or by infra-structural resources. By using five obvious instances in each case as the reference subset, a maximum resolution topology is derived to classify all 47 countries under study. As a result, 27 are classified as policy-dominant, and 20 as resource-dominant for their constraints in the investment climate.

The Global Diffusion of the Internet (GDI) Project, initiated by The Mosaic Group in 1997, focuses on measuring and analyzing the growth of the Internet throughout the world. Using a well-defined analytic framework (Wolcott *et al.*, 2001) with six dimen-

sions, nearly 30 countries have been studied. The six dimensions fall into a supply-demand dichotomy: with *Organizational Infrastructure*, *Connectivity Infrastructure*, and *Geographic Dispersion* being the supply dimensions, while *Pervasiveness*, *Sectoral Absorption*, and *Sophistication of Use* are the demand dimensions. Using data from 24 GDI cases for the year 1999, a maximum resolution topology is derived to classify countries according to whether their adoption of the Internet should be considered supply-side or demand side dominant at any particular stage of development. Initial international, inter-regional, as well as inter-temporal comparisons indicate that the MRT model can indeed shed light on subtle differences.

6. Discussion

We presented an optimization approach to derive an aggregate value function for the classification of multi-attribute dichotomies. Motivated by the Star Plot Method to visualize high-dimensional data, the area spanned by the attributes of a data instance suggests an aggregate measure in the form of the sum of products of adjacent pairs. Optimizing over the pairing of attributes as well as the weight on their products results in a maximum resolution topology with respect to a given subset of pre-classified dichotomies. Future work includes empirical studies and application of the approach in diverse fields, such as consumer market research, diffusion of innovation and technology, dynamics of online auction markets, and medical diagnostics. Theoretical results and refinements are also expected from further analysis of data normalization and harmonization, choice of spline functions, the process of selecting the reference subset, and alternatives to the constraints of bounded variations on the angles in the optimal topology.

Acknowledgment

Santosh Bukitgar and Christina Yuen provided able assistance at critical junctures of this project.

References

Ankerst, M., M. Ester and H.P. Kriegel (2000). Towards an effective cooperation of the computer and the user for classification. In *Proc. ACM SIGKDD Int. Conf. on Knowledge Discovery & Data Mining* (*KDD 2000*), Boston, MA. pp. 179–188.

Bohanec, M., and B. Zupan (2001). Integrating decision support and data mining by hierarchical multi-attribute decision models. In C. Giraud-Carrier, N. Lavrač, S. Moyle, B. Kavšek (Eds.), *IDDM-2001*: *ECML/PKDD-2001 Workshop Integrating Aspects of Data Mining, Decision Support and Meta-Learning*, Freiburg, pp. 25–36.

Chambers, J., W. Cleveland, B. Kleiner and P. Tukey (1983). *Graphical Methods for Data Analysis*. Wadsworth Press, Belmont, CA.

Chen, M.S., J. Han, P.S. Yu (1996). Data mining: an overview from a database perspective. *IEEE Transactions on Knowledge and Data Engineering*, **8**(6), 866–883.

Chu, S.C.K. (2001). A goal programming model for crew duties generation. *International Journal of Multi-criteria Decision Analysis*, **10**(3), 143–151.

Cline, A.K. (1974). Scalar- and planer-valued curve fitting using splines under tension. *Communications of the Association for Computing Machinery*, **17**, 218–220.

Dyer, J.S., and R.K. Sarin (1979). Measurable multi-attribute value functions. *Operations Research*, **27**, 810–822.

Ho, J.K. (2004). Topological analysis of online auction markets. *International Journal of Electronic Markets*, **14**(3), 202–213.

Hoffman, P., and G. Grinstein (2001). A survey of visualizations for high-dimensional data mining. In U. Fayyad *et al.* (Eds.), *Information Visualization in Data Mining and Knowledge Discovery*. Morgan Kaufmann. pp. 47–82.

Internet Legal Research Group (ILRG) (1996). Cost-benefit analysis of American law schools. `http://www.ilrg.com/schools/analysis/` (accessed February 14, 2005).

Johnson, C.A. (2000). Pricing gets personal. *The Forrester Report*, April 2000.

Keeney, R.L., and H. Raiffa (1976). *Decisions with Multiple Objectives*: *Preferences and Value Tradeoffs*. Wiley, New York.

Keim, D.A. (2002). Information visualization and visual data mining. *IEEE Transactions on Visualization and Computer Graphics*, **7**(1), 100–107.

Kirkwood, C.W., and R.K. Sarin (1980). Preference conditions for multi-attribute value functions. *Operations Research*, **28**, 225–232.

McGue, M. (1997). A behavioral-genetic perspective on children of alcoholics. *Alcohol Health and Research World*, **21**(3), 210–217.

Patel, A. (2003). Service use profiles of patients with schizophrenia or schizoaffective disorder across 15 countries. *Mental Health Research Review*, **9**, 24–26.

Poyhonen, M., and R.P. Hamalainen (2001). On the convergence of multi-attribute weighting methods. *European Journal of Operational Research*, **129**(3), 569–585.

Riesz, P.C. (1978). Price versus quality in the market place, 1961–1975. *Journal of Retailing*, **54**, 15–28.

Song, M., C.M. Breneman, J. Bi, N. Sukumar, K.P. Bennett, S. Cramer, N. Tugcu (2002). Prediction of protein retention times in anion-exchange chromatography systems using support vector regression. *Journal of Chemical Information and Computer Sciences*, **42**(6), 1347–1357.

Tukey, J.W. (1977). *Exploratory Data Analysis*. Addison-Wesley, Reading, MA.

Wolcott, P., L. Press, W. McHenry, S.E. Goodman and W. Foster (2001). A framework for assessing the global diffusion of the Internet. *Journal of the AIS*, **2**, article 6.
Available online: `http://mosaic.unomaha.edu/2001_GDI_Framework.htm` (accessed on June 6, 2005).

World Bank, The. (2004). *World Development Report 2005*: *A Better Investment Climate for Everyone*. Oxford University Press, New York.

J.K. Ho is a professor of information and decision sciences in the College of Business Administration at the University of Illinois at Chicago. He graduated from Columbia University in 1970 and obtained his PhD from Stanford University in 1974. His recent work focuses on parallel natural language generation in multilingual online communication, and topological analysis of online auction markets.

S.C.K. Chu is a faculty member in the Department of Mathematics at The University of Hong Kong. He graduated from Cornell University in 1973 and obtained his EngScD from Columbia University in 1978. His main area of interest is applied operations research. Along with the co-author, his recent work embarks on the optimization modeling aspect of topological analysis of online auction markets.

Maksimalios skiriamosios galios topologija daugelio atributų dichotomijose

James K. HO, Sydney C.K. CHU

Siūlomas optimizacinis modelis skirtas maksimizuoti dichotominių sprendimų skiriamąją galią, kai veikia daug skirtingų faktorių. Optimizavimas vykdomas pagal šių faktorių svorius lyginimo dichotomijų atžvilgiu. Pateikiami pavyzdžiai.