COST UNCERTAINTY MODELING FOR DESIGN-ANALYTIC CONCEPT EVALUATION

Karan Banga  
Mechanical Engineering Dept.  
Missouri S&T  
Rolla, MO, 65409  
kbcgf@mst.edu

Shun Takai  
Interdisciplinary Engineering Dept.  
Missouri S&T  
Rolla, MO, 65409-0210  
takais@mst.edu

ABSTRACT
During the product development stage, a number of decision making steps must be taken under uncertainty; one such step is selecting the ideal product concept. The cost of a concept, i.e., the cost of the final product developed from a concept, is a key factor influencing the choice of an ideal concept. This paper proposes a case based reasoning (CBR) approach to model beliefs about the uncertainty of the cost of a concept. The CBR approach creates a data base (or knowledge base) containing past cases, defines a new case, retrieves cases similar to the new case, adapts the solution of the retrieved cases to the new case, and stores the new case with its solution in the knowledge base. This paper illustrates the proposed approach using automobiles as an example.

KEYWORDS: cost, concept, clustering, distribution, discretization

1. INTRODUCTION
Product development involves a number of decision-making steps that must be taken under uncertainty, including selection of an ideal product concept. Factors influencing the choice of a concept include market size, market share, and concept cost (cost of the final product developed from a concept). Of all these factors, cost is the most critical. Two approaches have been proposed to determine the cost of a new concept, the bottom-up approach and the top-down approach. The bottom-up approach estimates product cost by adding costs associated with various product attributes and processes. This estimate is produced by adding part costs and assembly costs calculated from detailed product information such as the bill of information and the design specifications [1]. This approach requires detailed design and manufacturing process information, and since these processes are treated as uncertainties in the concept selection stage, this approach may not be useful in modeling the cost of a concept.

The top-down approach estimates product cost from product features and does not necessarily require design and manufacturing process information. Examples of top-down approaches are regression analysis and CBR. The CBR approach creates a data-base containing past cases, defines a new case, retrieves cases similar to the new case, adapts the solution of the retrieved cases to the new case, and stores the new case with its solution in the knowledge base [2]. It has been used in the past in various fields. Shepperd and Schofield [3] and Angelis and Stamelos [4] have used CBR to estimate the cost of new software projects. The number of projects retrieved was relatively small, i.e., ranging from one to three [3, 4], and the focus of the CBR approach was on calculating point estimates rather than constructing cost distributions [3, 4]. Finally, the cost of a new project was estimated by the mean of the costs of the retrieved projects [3, 4]. Jeffery et al. [5] compared the differences in accuracy between ordinary least squares regression and analogy based estimation using data from multiple companies as well as company-specific data.

The CBR approach has also been used in design problems. Bardasz and Zeid [6] have used it to solve mechanical design problems. Roderman and Tsatsoulis [7] have proposed the used of pumper apparatus novice design assistant (PANDA), a case-based design system, to assist firefighters who wish to design their pumper engines. Maier and Zhang [8] have proposed a case-based design process model, CADSYN, to solve new design problems. The CBR approach has also been used to estimate costs of construction projects. Kim et al. [9] and An et al. [10] have discussed such uses.

Although CBR has been used in the fields mentioned above, it has not been used to estimate the cost of a product. This paper proposes a CBR approach to model beliefs about the uncertainty of the cost of a concept (or product). Current methods rely heavily on point estimates rather than constructing distributions, and with as few as three projects retrieved (as in the case of [3] and [4]), no distribution can be constructed. Unlike the approaches mentioned above, that presented here constructs a cost distribution for the products retrieved from the data base to estimate the cost of a concept.
2. METHODOLOGY

When a company selects a new product concept for launching in the market, the cost of the concept is the most important factor in the decision. This paper proposes the use of CBR in estimating the cost of a new concept, using automobiles as an example. This approach involves four steps: (1) constructing a knowledge-base of all the existing products available in the market, (2) defining a new product concept according to our needs, (3) retrieving from the data base, the products most similar to the concept, and (4) fitting a distribution curve for the required property. The complete process is illustrated in figure 1.

![Figure 1: CBR Process](image)

2.1 Knowledge-Base construction

The first step in any CBR approach is to construct a knowledge base of all the products available in the market. The products, together with their attributes, are listed in the knowledge base. The attributes may be quantitative or categorical, and the attributes selected are crucial to the analysis. One way of selecting the attributes is to conduct a market survey and incorporate customer requirements and needs. Another approach would rely on previous experience and include the attributes that have proved most important over a period of time.

2.2 Product concept definition

The next step in the CBR approach is to define a new product concept with the selected attributes. The attributes for the new concept are chosen, to a large extent, based on customer preferences. They may also depend on the availability of attribute data in the knowledge base.

2.3 Product retrieval

Perhaps the most important step in CBR analysis is to retrieve from the knowledge-base, products with attributes similar to those of the concept. A number of methods have been proposed for this retrieval; this paper relies on use of hierarchical clustering analysis. Hierarchical clustering analysis could be broken down into a number of steps: (1) Data matrix creation: If the knowledge base contains i number of products and j number of attributes, then the data matrix would have i+1 rows and j columns, with the additional row accommodating the new concept. (2) Distance matrix creation: The distance matrix is created from the data matrix by calculating Euclidian distance between the concept and each product, and between each pair of products. (3) Hierarchical clustering: Finally, hierarchical clustering is applied to the distance matrix to obtain products similar to the concept. Hierarchical clustering is used to generate tree-figures, also called as dendrograms, based on the distances calculated in the distance matrix. The products are grouped at various levels based on the distances between them. Products separated by smaller distances are grouped at a lower level because they are similar, (i.e., they have similar information). All these steps can be performed using commercially available software discussed below.

2.4 Cost histogram construction and distribution fitting

The next step in CBR analysis is distribution fitting. Once products which are similar to the concept are retrieved, a distribution must be fitted to the desired property. Generally, histograms of the desired property are constructed first, and a distribution is then fitted to them. Normal distribution is used most often, but other distributions may also be used. The present analysis uses normal distribution to explain the distribution pattern.

2.5 Discretization

The last step in the analysis is to discretize the continuous distribution. Here, the extended Pearson-Tukey [11] method is used. This is a three point approximation method to calculate three percentile values corresponding to probability values: 0.05, 0.50 and 0.95. For example, probability 0.05 and cost \( C_{0.05} \) may be selected such that \( P(\text{Cost} \leq C_{0.05}) = 0.05 \). This number \( C_{0.05} \) is called 0.05 fractile of the distribution. These three percentile values are then assigned probability values of 0.185, 0.630 and 0.185 for 0.05 fractile, 0.50 fractile and 0.95 fractile respectively. The complete process is shown in Fig. 2.
3. ILLUSTRATIVE EXAMPLE EXPLAINING CASE BASED REASONING

In order to see the application of CBR, we look at an illustrative example using automobiles to explain the procedure. For the sake of simplicity, this example uses a data base which is condensed, in terms of the number, both of cars and their attributes. Section 3.4, however, presents the cost distribution curve for the concept using the original data set. The steps involved in the analysis are described in detail in the following sections.

3.1 Knowledge-base construction

The original knowledge base originally comprised a pool of various types of cars currently available in U.S., with their attributes. The number of cars in the pool was close to 300, and the number of attributes was close to 80. The attributes in the knowledge base included type of car, number of cylinders, engine capacity, number of side airbags, acceleration, braking, overall mileage (in mpg), roadside aid and many more. The data on cars and their attributes were gathered primarily from the internet, and the data on costs were taken from the annual reports of individual automobile companies. Table 1 shows a part of the original data-base.

<table>
<thead>
<tr>
<th>Automobile type</th>
<th>overall mileage, mpg</th>
<th>Cost($)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Porsche Boxster convertible</td>
<td>22</td>
<td>21554.26</td>
</tr>
<tr>
<td>Toyota RAV4 4-cyl. SUV</td>
<td>23</td>
<td>20994.05</td>
</tr>
<tr>
<td>Honda CR-V SUV</td>
<td>21</td>
<td>17483.00</td>
</tr>
<tr>
<td>Mitsubishi Lancer ES small cars</td>
<td>25</td>
<td>14332.37</td>
</tr>
<tr>
<td>Mercedes-Benz E320 sedans</td>
<td>29</td>
<td>53100.08</td>
</tr>
<tr>
<td>Chrysler Town &amp; Country Limited Minivans</td>
<td>17</td>
<td>37594.98</td>
</tr>
<tr>
<td>Toyota Prius Touring Wagons</td>
<td>42</td>
<td>22574.88</td>
</tr>
<tr>
<td>Dodge Dakota pickups</td>
<td>14</td>
<td>28052.69</td>
</tr>
<tr>
<td>Ford Mustang V8 Sporty</td>
<td>20</td>
<td>27413.44</td>
</tr>
</tbody>
</table>

Table 1: Part of the original data base

As shown in Table 1, the data base included cars of various types, e.g., convertibles, SUVs, small cars, and so on. To standardize costs, the attribute cost was calculated using Eq. (1):

\[ \text{Cost} = \text{price} \times \frac{\text{avg. cost}}{\text{avg. revenue}} \] (1)

The average cost and average revenue used in this equation were taken from annual reports of car companies. The attribute cost, though mentioned in the data base, was not used in the analysis.

3.2 Concept definition

The concept tested was an SUV with overall mileage of 40 mpg. The concept with its attributes is described in Table 2. The two attributes most crucial for the analysis are considered here: automobile type and overall mileage.

<table>
<thead>
<tr>
<th>Concept</th>
<th>Automobile type</th>
<th>overall mileage, mpg</th>
<th>Normalized overall mileage</th>
</tr>
</thead>
<tbody>
<tr>
<td>SUV</td>
<td>40</td>
<td>0.88</td>
<td></td>
</tr>
</tbody>
</table>

Table 2: Concept selection

3.3 Clustering Analysis

For this analysis, the original data-base was modified as shown in Table 3.

<table>
<thead>
<tr>
<th>SNo</th>
<th>convertibles</th>
<th>SUV</th>
<th>minivans</th>
<th>sedans</th>
<th>minivans</th>
<th>Wagons</th>
<th>pickups</th>
<th>Sporty</th>
<th>overall mileage, mpg</th>
<th>Normalized overall mileage</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>P1</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>40</td>
<td>0.88</td>
</tr>
<tr>
<td>2</td>
<td>Toyota RAV4 4-cyl</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>21</td>
<td>0.35</td>
</tr>
<tr>
<td>3</td>
<td>Honda CR-V</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>21</td>
<td>0.35</td>
</tr>
<tr>
<td>4</td>
<td>Mitsubishi Lancer ES</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>25</td>
<td>0.44</td>
</tr>
<tr>
<td>5</td>
<td>Mercedes-Benz E320</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>25</td>
<td>0.56</td>
</tr>
<tr>
<td>6</td>
<td>Chrysler Town &amp; Country Limited</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>17</td>
<td>0.31</td>
</tr>
<tr>
<td>7</td>
<td>Toyota Prius Touring</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>42</td>
<td>0.94</td>
</tr>
<tr>
<td>8</td>
<td>Dodge Dakota</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>14</td>
<td>0.12</td>
</tr>
<tr>
<td>9</td>
<td>Ford Mustang V8</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>20</td>
<td>0.39</td>
</tr>
</tbody>
</table>

Table 3: Modified data base

In Table 3, the cars are listed in rows and the attributes in columns. The first row contains the concept vehicle and the first column denotes a unique identifier for the automobiles. The first attribute, type of automobile has been broken down into convertibles, SUVs, small cars, sedans, minivans, wagons, pickups, and sports cars (sporty). The two attributes used here are different in nature; the first attribute, type of automobile is categorical in nature and hence 0's and 1's are sufficient to define it. The second attribute, overall mileage is quantitative and therefore, it has been normalized to negate the influence of the units used. In other words, each attribute value is normalized between 0 and 1 so that it has the same degree of influence as other attributes. Also, as mentioned above, the attribute cost was not used in this analysis. This attribute was used only to ascertain cost and only after the analysis was run and cars similar to the concept were obtained. The cost of the concept was the outcome of the overall analysis. The normalization technique used is shown in Equation (2).
\[ d_i = \frac{X_i - X_{\text{min}}}{X_{\text{max}} - X_{\text{min}}} \]  

(2)

Where \( d_i \) is the normalized attribute value, \( X_i \) is the original attribute value, and \( X_{\text{max}} \) and \( X_{\text{min}} \) are the maximum and minimum values across an attribute \( i \), where \( i \) varies from 1 to \( n \).

Overall mileage varied from 10 to 44 mpg; therefore, the minimum and maximum values used were 10 mpg and 44 mpg, respectively.

Once the database was ready for the analysis, the computer software, MATLAB was used to perform clustering analysis and retrieve products similar to the concept. Table 3 was used as an input. The resulting dendrogram is shown in fig. 1. The three most commonly used hierarchical clustering methods are: complete linkage method, average linkage method, and Ward’s method. This analysis used the complete linkage method, which calculates, element by element, the distances between two clusters and uses the maximum distance as the distance of two clusters. In the dendrogram, the height at which two cars, two clusters, or a car and a cluster are grouped together is the distance between them. Hence, similar cars are grouped at a lower level because they have smaller distances between them.

This distance is called Euclidian distance, \( \delta \), and between a new product \( p' \) and an existing product \( p \) is given by Equation (3).

\[ \delta(p, p') = \sqrt{\sum_{i=1}^{n} w_i (d_i - d_{i}')^2} \]  

(3)

Where \( d_i \) is the normalized attribute of product \( p \) (as explained above for Equation 2), \( d_{i}' \) is the normalized attribute of product \( p' \), and \( w_i \) is the weight of attribute \( i \), where \( i \) varies from 1 to \( n \). In this case, unweighed Euclidian distance was used, i.e. \( w_i = 1 \) for \( i = 1 \) to \( n \).

Once the dendrogram is created, cars similar to the concept must be selected. This study defines as similar those cars that are one linkage below the concept. Figure 3 shows the dendrogram obtained from the analysis.

As shown in fig. 3, cars P2 and P3 are clustered one linkage below Concept C1 and are thus considered similar to the concept.

3.4 Cost distribution

Once similar cars have been identified, the next step is to fit a distribution. Here, a normal distribution is used. Usually, the histogram is constructed first, and a distribution is then fitted for the desired property. In this case, the desired property is the cost of the concept; therefore, the costs of the retrieved cars are used to construct the histogram and then to fit a distribution. The cost distribution curve for the concept using the complete data set is shown in fig. 4.
3.5 Discretization

Once the distribution is fitted, the final step is to discretize the distribution or find the point estimates from the distribution. This has been accomplished here using the extended Pearson-Tukey method. Probability distribution tables commonly available for normal distribution were used and values were determined corresponding to the probability values of 0.05, 0.50 and 0.95. The results can be summarized as follows: 
\[ C_{0.05} = 16691, \quad C_{0.50} = 33884, \quad C_{0.95} = 51078 \]

Hence, the distribution in Fig. 3 can be discretized approximately using three values: 16691, 33884 and 51078 with the associated probabilities 0.185, 0.63, 0.185 respectively, as shown in Fig. 5

![Figure 5: Final Three-Point Discretization](image)

4. Conclusion and Future Work

The work applied the CBR approach to model beliefs about the uncertainty of the cost of a concept. A knowledge base was constructed of all the cars available on the market, and a new car concept was defined taking type and overall mileage as the two attributes. The cars most similar to the concept were retrieved from the knowledge base, and a distribution curve was fitted to the costs of the retrieved cars. Finally, the distribution was discretized to obtain a three approximation.

Future work will adjust retrieved costs using a regression model before fitting a cost distribution. Such work may also include leave-one-out cross-validation of the costs and evaluation of the various approaches.

ACKNOWLEDGEMENTS

We would like to thank the Intelligent Systems Center and the Interdisciplinary Engineering Department at Missouri University of Science and Technology for supporting this research.

5. REFERENCES