Res Eng Design (2009) 20:255–264
DOI 10.1007/s00163-009-0067-4

ORIGINAL PAPER

A case-based reasoning approach toward developing a belief about the cost of concept

Shun Takai

Received: 4 November 2008 / Accepted: 17 March 2009 / Published online: 15 April 2009

Abstract It is generally acknowledged that product development involves a sequence of decision making under uncertainty, including setting target requirements for a new product, selecting product concept, and developing conceptual and detailed design of a chosen concept. To select a product concept, engineers need to assess the uncertainty of a future market share, market size, and a cost of concept (cost of the final product developed from a concept). This paper proposes a case-based reasoning (CBR) approach to model beliefs about the uncertainty of a cost of concept. The proposed CBR approach consists of storing information about various products in a knowledge-base, defining a new product concept, retrieving a cluster of products in the knowledge-base that are highly similar to the concept, and adapting the cost of the retrieved product to construct a distribution of the cost of concept. This paper illustrates the proposed approach using printers as an example.

Keywords Cost · Concept · Clustering · Distribution

1 Introduction

Product development typically consists of a sequence of activities and stages in Fig. 1. Concept selection is one of the most important stages because the product development efforts after this stage will be based on the selected concept. Primary functions, technologies, and performance specifications of a concept are specified in the concept selection stage before deciding detail design and manufacturing processes (i.e., before deciding design specifications). An example of a performance specification is printing speed in the case of a printer and examples of design specifications are weight, dimensions, product architecture, material used, and manufacturing processes.

Concept selection also involves large degrees of uncertainties because engineers need to make decisions when information is incomplete, that is, without knowing the future market size, market share, and the cost. To estimate the profitability of each concept, revenue is calculated by multiplying the market size, market share, and price, and profit is calculated by subtracting the cost from the revenue. Market size may be forecasted from the past data and market share may be estimated using a market survey. On the other hand, cost needs to be estimated by engineers.

Product cost may be estimated using bottom-up or top-down approach. In the bottom-up approach, a product cost is estimated by adding costs associated with different product attributes. The bottom-up approach includes cost models that estimate product cost by adding part costs and assembly costs calculated from detailed product information such as bill of material (BOM) and design specifications (Ulrich and Eppinger 2004; Otto and Wood 2001; Pahl and Beitz 1996; Dewhuest and Boothroyd 1988); activity-based costing (ABC), which estimates product cost from BOM and activities necessary to manufacture these parts (Cooper and Kaplan 1987); and feature-based costing, which estimates product cost by mapping cost to product features (Brimson 1998). These bottom-up approaches require detailed product design and manufacturing process information; therefore, they may not be used in estimating the cost of the final product developed from a concept (a cost of concept,

S. Takai (✉)
Department of Interdisciplinary Engineering,
Missouri University of Science and Technology,
101 Interdisciplinary Engineering Building,
1215 N. Pine St, Rolla, MO 65409-0210, USA
e-mail: takais@mst.edu

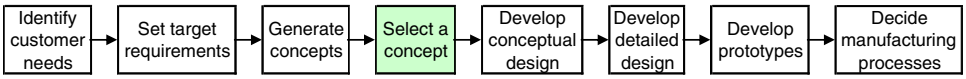

Fig. 1 Product development process

hereafter). Design and manufacturing process are treated as the uncertainties in the concept selection stage.

Different from bottom-up approaches, top-down approach estimates product costs from product features and do not necessarily require detailed design and manufacturing process information. Examples of a top-down approach are regression analysis or parametric cost estimation (Hamaker 1995; Wyskida 1995), a curve fitting approach that fits a curve to historical data and estimates the cost of a product from the curve (Pugh 1990), an artificial neural networks application (Seo et al. 2002), and case-based reasoning (CBR) applications.

The importance of estimating the cost of a product and its parts in the conceptual design stage has been emphasized repeatedly (Mileham et al. 1993; Newnes et al. 2008; Mauchand et al. 2008). Parametric cost estimation, which identify cost estimating relations (CERs) between cost and cost drivers using regression analysis, has been widely used in estimating costs of government projects (Joint Industry/Government Parametric Cost Estimating Initiative 1999), of parts (Watson et al. 2006; Hicks et al. 2002), and of products (Curran et al. 2005; Roy et al. 2001). For example, parametric approach has been used to model manufacturing and life cycle cost in the conceptual design stage from design parameters (total weight, part count, etc.) estimated in the conceptual design stage (Curran et al. 2005).

Compared to parametric cost estimation, CBR is a relatively new approach used in estimating costs of new software projects (Shepperd and Schofield 1997; Auer et al. 2006; Angelis and Stamelos 2000; Mendes et al. 2003; Jeffery et al. 2000) and costs of construction projects (Kim et al. 2004; An et al. 2007). CBR process consists of *storing* cases (past products) in the knowledge-base, *defining* a new problem (product concept), *retrieving* cases that are similar to the new problem, *adapting* the solution of the retrieved cases (cost past products) to the new problem (estimating the cost of concept), and storing the solved problem and its solution in the knowledge-base (Kolodner 1993).

In CBR cost estimation approaches, a "case" that is a historical project p stored in the knowledge-base is described by its cost c (or software development effort as a proximity of cost) and a list (vector) of features $\{d_1, d_2, \ldots, d_n\}$. Once the features $\{d_1', d_2', \ldots, d_n'\}$ of the new project p' are specified, the projects that are similar to the new project are retrieved from the knowledge-base. In CBR approaches, engineers need to decide the metric for similarity (or distance) between a new project and a historical project, the relative importance of a feature in calculating the distance, the number of closest cases to retrieve, and the adaptation methodology to use. Although any distance measure may be used, the most popular distance metric between a new project p' and a historical project p is the Euclidean distance δ in Eq. 1. The difference of each feature i, $d_i - d_i'$, is normalized by dividing the difference by the maximum of d_i, $d_{i\max}$. The weight of each feature w_i is equal to one for the case of an unweighted Euclidean distance.

$$\delta(p, p') = \sqrt{\sum_{i=1}^{n} wi\left(\frac{d_i - d_i'}{d_{i\max}}\right)^2}. \quad (1)$$

In the case of a weighted Euclidean distance, the analytic hierarchy process (An et al. 2007), gradient descent method (Kim et al. 2004), and extensive search from a set of possible weights (Shepperd and Schofield 1997; Auer et al. 2006) have been proposed as potential approaches for specifying feature weights. The number of retrieved projects in the CBR cost estimation approaches is relatively small—typically up to three (Shepperd and Schofield 1997; Mendes et al., 2003; Jeffery et al., 2000; Kim et al. 2004; An et al. 2007) and the focus of the CBR approaches is on improving the accuracy of point estimates rather than constructing cost distributions. The cost of a new project is estimated by the mean or median of the costs of the retrieved projects (Shepperd and Schofield 1997; Mendes et al. 2003; Jeffery et al. 2000; Kim et al. 2004; An et al. 2007). The cost is not adjusted before calculating the mean or median in most of the CBR applications, except for the linear-adjustment studied in Jeffery et al. (2000). Although CBR has been widely used in many design problems (Bardasz and Zeid 1991, 1993; Roderman and Tsatsoulis 1993; Maher and Zhang 1993; Shiva Kumar and Krishnamoorthy 1995; Wood and Agogino 1996; Al-Shihabi and Zeid 1998; Rosenman 2000; Lee and Lee 2002), it has not been used for constructing cost distributions and for estimating the cost of concept.

To account for uncertainty and the lack of detailed design and manufacturing process information in the conceptual design stage, this paper proposes a CBR approach toward modeling beliefs about uncertainty in terms of distributions as well as calculating point estimates for the cost of concept. In particular, this paper applies a hierarchical clustering method to retrieve data that are as homogeneous as possible and that are highly similar to the concept from a heterogeneous knowledge-base, and use regression analysis to adapt the retrieved data to construct distributions. The rest of this paper consists of the following sections: Sect. 2 describes the proposed framework; Sect. 3 presents an illustrative example; Sect. 4 compares

the accuracy of point estimates for three different cost adaptation (adjustment) approaches; and Sect. 5 contains the conclusion and future work.

2 Proposed framework

Using printers as an example, Fig. 2 schematically illustrates the framework proposed in this paper toward constructing distributions of a cost of concept. First, engineers store information about various printers in the knowledge-base. Once a new printer concept is defined, engineers search and retrieve a cluster of printers in the knowledge-base that are highly similar to the concept. The costs of retrieved printers are adapted to model beliefs about the cost of concept in a form of distribution. Because the design (design specifications) of the product has not been decided in the concept selection stage, the cost of concept is a probabilistic assessment of a future condition including design.

Figure 3 illustrates the six-step procedure of the proposed CBR framework: knowledge-base construction (Step 1), concept definition (Step 2), case retrieval (Step 3), cost relevant specification identification (Step 4), cost adjustment (Step 5), and distribution construction (Step 6).

Step 1 Knowledge-base construction The first step is to construct a knowledge-base by storing information of various products. Product information includes costs, primary functions, technologies, requirements, and specifications. Requirements refer to metrics and specifications refer to specific target values of the requirements. For example, "fuel efficiency" is a requirement and "30 miles per gallon" is a specification. Because products in the knowledge-base are the final products, product specifications consist of *performance specifications*, *design specifications*, as well as other specifications, including warranty and industry standards. In the case of a printer, printing speed is an example of performance specifications, and weight and dimensions are examples of design specifications.

Step 2 Concept definition The second step is to define a concept by its primary functions, technologies, and performance specifications. Performance specifications may be specified by mapping customer requirements to product requirements using quality function deployment (QFD) and by setting a target value of product requirements (Hauser and Clausing 1988; Clausing 1993). Because detailed design is not developed at this stage, there is no design specification for the concept.

Step 3 Case retrieval The third step is to retrieve products in the knowledge-base that have similar product information (primary functions, technologies, and specifications) as the concept. Hierarchical clustering method (Mardia et al. 2000) is used to retrieve as many products as possible that have information similar to the concept while information among retrieved products is as homogeneous as possible. Information is considered homogeneous for each of the information categories, if all products in the cluster have the information or none of the products have the information. To assess homogeneity, a data matrix is constructed for the concept and each product by entering 1 if the product information is available and 0 if otherwise in each product information category. If there are m number of products and n number of information categories in the knowledge-base, the data matrix, which consists of 0 and 1, is $m + 1$ rows (m products and a concept) by n columns. Then an $m + 1$ rows by $m + 1$ columns distance matrix is constructed from the data matrix by calculating the Euclidean distance between the concept and each product and between each pair of products. Finally, hierarchical clustering is applied to this distance matrix to group products with similar product information. To identify the optimum number of clusters, this paper compares the clusters obtained by three hierarchical clustering methods—complete linkage method, average linkage method, and Ward's method—and chooses the clusters that are common to all three methods, i.e., the clusters robust to the hierarchical clustering methods. Products similar to the concept are retrieved from the cluster that is grouped together with the concept in the hierarchical clustering. Hierarchical

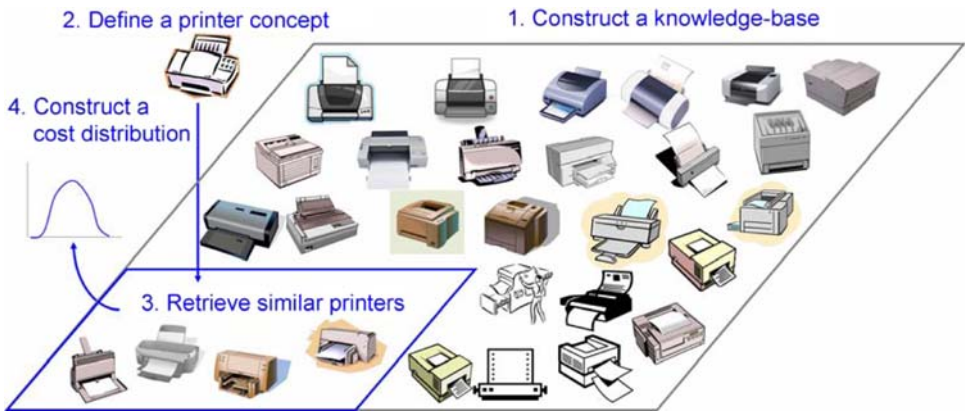

Fig. 2 Framework

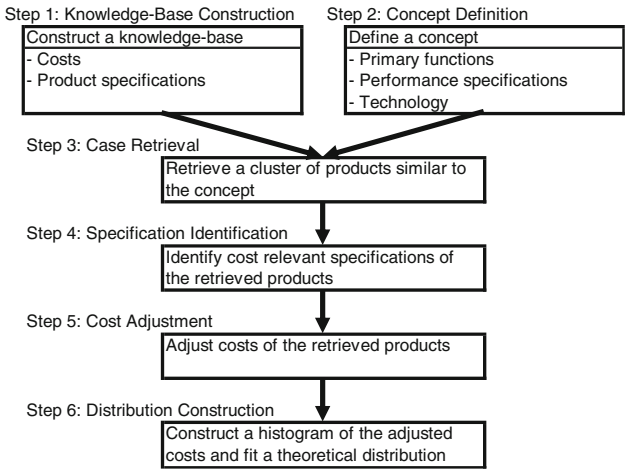

Fig. 3 Procedure

clustering has been used to group similar cases in the knowledge-base before retrieving the most similar case from the group (Reich and Kapeliuk 2004). In the proposed approach all the cases in the group similar to the concept is retrieved to estimate costs and to construct cost distributions.

Step 4 Specification identification The fourth step is to identify cost relevant performance specifications of the retrieved products. In this step, multiple linear regression analysis (Neter et al. 1996) is performed. In the regression analysis, cost (dependent variable) is regressed on performance specifications of the retrieved products (independent variables).

Step 5 Cost adjustment The fifth step is to adjust the costs of retrieved products using the regression model obtained in Step 4. The cost is adjusted parallel to the regression line (or surface, if there is more than one numeric cost relevant specification) for the difference between cost relevant specifications of the concept and those of the retrieved products.

Step 6 Distribution construction The final step is to construct a histogram of the adjusted costs and construct a distribution by fitting a theoretical distribution, e.g., a normal distribution. The fitted distribution is the distribution of the cost of concept.

3 Illustrative example

This section illustrates the proposed approach using printers as an example.

3.1 Construct knowledge-base

The knowledge-base is constructed by collecting product information (costs, functions, technologies, and specifications) of 69 printers from a manufacturer's website. To avoid any bias due to human judgment, all the information has been stored in the knowledge-base. There are 102 product information categories. Among these categories, 47 are numeric and 55 are categorical. The costs of these printers are obtained by an approach similar to top-down target costing (Ulrich and Eppinger 2004), in which a cost is estimated by subtracting a profit margin from a price. The profit margin is estimated by averaging gross profit margins in the manufacturer's annual 10-K financial reports for the years 2001 through 2005. This number is verified by the profit margin in Ulrich and Eppinger (2004).

3.2 Define concept

The concept of a printer is defined by its primary functions and technologies listed in Table 1, and performance specifications partially listed in Table 2.

3.3 Retrieve products similar to the concept

To retrieve printers similar to the concept, engineers first construct a data matrix that represents the information available for both the concept and products in the knowledge-base. Different product types (e.g., mono laser, color laser, and inkjet printers in the case of printers) may have different sets as well as a common set of parameters (performance and deign specification categories). For example, black printing speed is a performance specification common for both mono and color laser printers. On the other hand, color printing speed is a performance specification only for color laser printer.

The heterogeneity of parameters is summarized in the data matrix in which 1 is used if a concept or a case (a product in the knowledge-base) is described by a performance parameter and 0 if otherwise. Figure 4 is a portion of the data matrix in which C1 is the concept, P1, P2, and so forth are printers in the knowledge-base, and I1, I2, and so forth are product information categories. For the concept and for each printer, 1 is entered if information is available and 0 if otherwise for each information category.

Table 1 Primary function and technology

Function and technology	Specifications
Function	Color printing
	Color scanning
	Color copying with PC
Print technology	Thermal inkjet
Color technology	Four color inkjet (cyan, magenta, yellow, black)
Scan technology	CIS with 48 bit depth

Table 2 Representative performance specifications (partial list)

Requirements	Specifications
Maximum print speed (draft, black): up to (ppm)	17
Maximum print speed (draft, color): up to (ppm)	9
Maximum print speed (normal, black): up to (ppm)	10
Maximum print speed (normal, color): up to (ppm)	3
Maximum copy speed (draft, black): up to (cpm)	15
Maximum copy speed (draft, color): up to (cpm)	6
Print resolution, black: up to (dpi × dpi)	2,400 × 1,200
Print resolution, color: up to (dpi × dpi)	4,800 × 1,200
Scan resolution, enhanced: up to (dpi × dpi)	9,600
Scan resolution, optical (dpi × dpi)	600 × 1,200
Print noise level, operating (<dBA)	44
Copy noise level, operating (<dBA)	44
Scan noise level, operating (<dBA)	38

From this data matrix, the distance matrix in Fig. 5 is constructed by calculating Euclidean distances between printers using Eq. (1). The weights are set equal (i.e., $w_i = 1$ for $i = 1,\ldots, 102$) for all 102 information categories (I1, I2,…, I102), and $d_{i\max} = 1$ and $d_i - d_i' = 0$ or 1 because availability of information are expressed by either 0 or 1 in Fig. 4. Comparing C1 and P1, because there are ten information categories in which P1 has data and C1 does not (not shown in Fig. 4), the distance between C1 and P1 is $\sqrt{10} = 3.16$. This distance is shown in the entry in the first row and the second column (or that of the second row and the first column) in the distance matrix.

Then three hierarchical clustering methods—complete linkage method, average linkage method and Ward's method—are used in order to generate three hierarchical clustering trees (dendrograms) of the concept and printers from the distance matrix (Fig. 13 in Appendix A). The complete linkage method calculates, element by element, distances between two clusters and uses the maximum distance as the distance of two clusters. Average distance method also calculates, element by element, distances of two clusters but uses the average distance as the distance of two clusters. Ward's method groups elements so that within cluster variances are minimized.

In the dendrogram, the height at which two printers, two clusters, or a printer and a cluster are grouped together is the distance between them. Thus, similar printers (printers with similar information) are grouped together at a lower level because they have smaller distances. For example, the set of available product information is the same for printers P32 and P33. Thus, the distance between these printers is

Fig. 4 Data matrix (portion)

	I1	I2	I3	I4	I5	I6	I7	I8	I9	I10	I11	I12	I13	I14	I15
C1	1	1	1	1	1	1	1	1	1	1	1	1	0	0	0
P1	1	1	1	1	1	1	1	1	1	1	1	1	0	0	0
P2	1	1	1	1	1	1	1	1	1	1	1	1	0	0	0
P3	1	1	1	1	1	1	1	1	1	1	1	1	0	0	0
P4	1	1	1	1	1	1	1	1	1	1	1	1	0	0	0
P5	1	1	1	1	1	1	1	1	1	1	1	1	0	0	0
P6	1	1	1	1	1	1	1	1	1	1	1	1	0	0	0
P7	1	1	1	1	1	1	1	1	1	1	1	1	0	0	0
P8	1	1	1	1	1	1	1	1	1	1	1	1	0	0	0
P9	1	1	1	1	1	1	1	1	1	1	1	1	0	0	0
P10	1	1	0	1	0	0	0	0	0	0	0	0	1	1	1
P11	1	1	0	1	0	0	0	0	0	0	0	0	1	1	1
P12	1	1	0	1	0	0	0	0	0	0	0	0	1	1	1
P13	1	1	0	1	0	0	0	0	0	0	0	0	1	1	1
P14	1	1	0	1	0	0	0	0	0	0	0	0	1	1	1

Fig. 5 Distance matrix (portion)

	C1	P1	P2	P3	P4	P5	P6	P7	P8	P9	P10	P11	P12	P13	P14
C1	0	3.16	3.46	3.32	3.74	3.74	4	4.24	3.87	4.12	4.36	7.62	8.06	8.12	8.12
P1	3.16	0	1.41	1	2	2	2.45	2.83	2.24	2.65	3	7.21	7.68	7.75	7.75
P2	3.46	1.41	0	1	1.41	1.41	2.83	2.45	1.73	2.24	2.65	7.35	7.81	7.87	7.87
P3	3.32	1	1	0	1.73	1.73	2.65	2.65	2	2.45	2.83	7.28	7.75	7.81	7.81
P4	3.74	2	1.41	1.73	0	1.41	2.83	2	1	2.24	2.24	7.35	7.68	7.75	7.75
P5	3.74	2	1.41	1.73	1.41	0	2.83	2	1	1.73	2.24	7.48	7.81	7.87	7.87
P6	4	2.45	2.83	2.65	2.83	2.83	0	2.45	3	2.24	2.65	7.62	7.94	8	8
P7	4.24	2.83	2.45	2.65	2	2	2.45	0	1.73	1	1	7.62	7.94	8	8
P8	3.87	2.24	1.73	2	1	1	3	1.73	0	2	2	7.42	7.75	7.81	7.81
P9	4.12	2.65	2.24	2.45	2.24	1.73	2.24	1	2	0	1.41	7.68	8	8.06	8.06
P10	4.36	3	2.65	2.83	2.24	2.24	2.65	1	2	1.41	0	7.55	8	7.94	7.94
P11	7.62	7.21	7.35	7.28	7.35	7.48	7.62	7.62	7.42	7.68	7.55	0	3	2.83	2.83
P12	8.06	7.68	7.81	7.75	7.68	7.81	7.94	7.94	7.75	8	8	3	0	1	1
P13	8.12	7.75	7.87	7.81	7.75	7.87	8	8	7.81	8.06	7.94	2.83	1	0	0
P14	8.12	7.75	7.87	7.81	7.75	7.87	8	8	7.81	8.06	7.94	2.83	1	0	0

zero and they are grouped together at the lowest level, i.e., height = 0. The clusters of printers are identified by cutting the dendrogram at an arbitrary height. For example, the horizontal lines in Fig. 13 cut the dendrogram that results in six clusters of printers (ignoring concept C1): A, B, C, D, E, and F. Printers in these clusters are P1–P10 in A, P11–P27 in B, P28–P39 in C, P40–P43 in D, P44–P64 in E, P65–P69 in F. The optimum set clusters are defined by the clusters that are common in three hierarchical clustering methods.

Figure 6 compares 2, 3, 4, 5, and 6 clusters obtained by the three methods. All three methods result in the same set of clusters in the case of two and six clusters (ignoring concept C1). The final choice of the number of clusters involves a tradeoff between the number of printers in each cluster and the homogeneity of product information among printers in each cluster. To adjust the costs of retrieved printers by applying regression analysis in the next step, it is desirable that the clusters contain a large number of printers and the information is highly homogeneous. Increasing the number of clusters improves homogeneity of information of printers in each cluster; however, it reduces the number of printers in each cluster. Six clusters are chosen so that the number of printers in each cluster is relatively large and the information is highly homogeneous. Printers similar to the concept C1 are those in the cluster grouped with C1. These are printers P1–P10 in cluster A, as illustrated in Fig. 13.

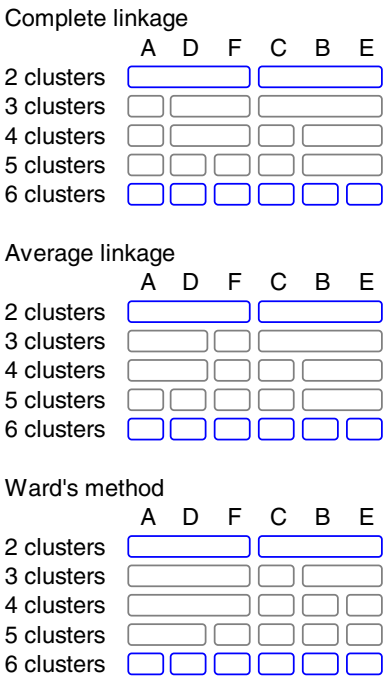

Fig. 6 Common clusters

3.4 Identify cost relevant requirements

Among 102 variables (information categories), variables are omitted if all printers in the chosen clusters have the same or no data, and if variables are redundant. Missing specification is replaced by 0 if it is a numeric specification and N/A if it is a categorical specification. After screening the variables, multiple linear regression analysis is performed to identify cost relevant specifications. Equation 2 is the regression model obtained from regression analysis. The intercept is approximately $3; however, it is not significantly different from 0. In Eq. 2, "Maximum Copy Speed (draft, black)" is the numerical variable for the copy speed (cpm), and "Resolution" is the categorical variable for the optical scan resolution. Resolution is 1 if the printer has a better resolution (1,200 × 4,800 dpi) and 0 if otherwise. Only one printer has a better resolution. The coefficient for the maximum copy speed is significant at a 5% level and that of resolution is significant at a 1% level. Adjusted R^2 is 0.9151.

$$\text{Cost} = 2.893 + 3.7 \times \text{Maximum Copy Speed} + 88.05 \times \text{Resolution} \quad (2)$$

Figure 7 illustrates costs of the ten printers and the corresponding regression line. Only nine data are shown in the figure because two printers have the same maximum copy speed and cost.

Figure 8 shows the normal probability plot of residuals (residuals plotted against their expected values under normality assumption). Reasonably straight relationships between observed residuals and theoretical values support that normal distribution is a reasonable assumption for the distribution of residuals. The coefficient of correlation 0.946 between observed and theoretical residuals supports the assumption of normally distributed residuals, because it is larger than the critical value of 0.918 for ten samples at an α risk level of 0.05 (Looney and Gulledge 1985; Neter et al. 1996).

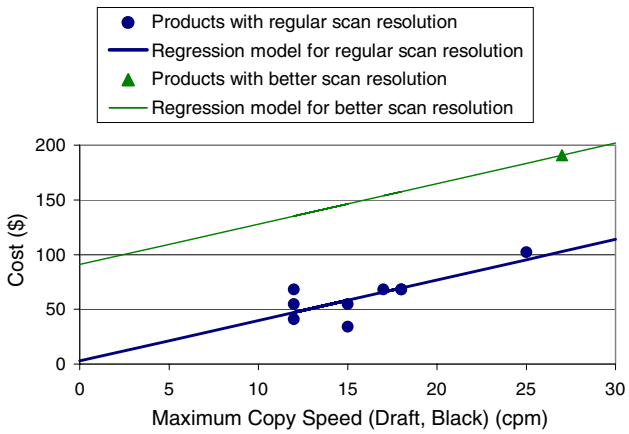

Fig. 7 Regression model

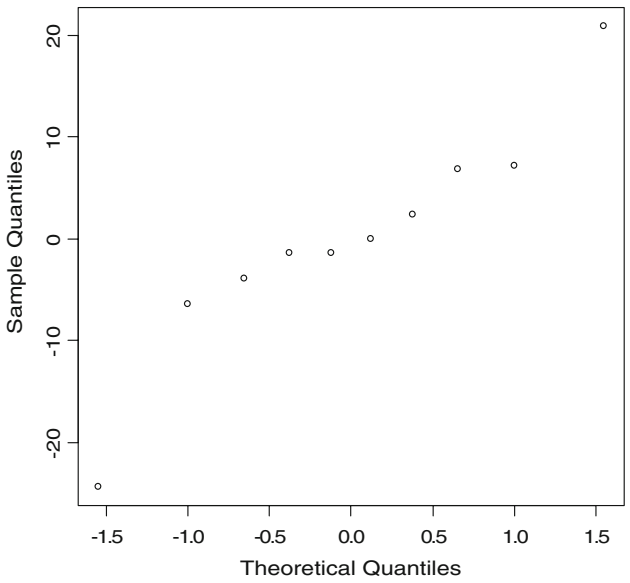

Fig. 8 Normal probability plot of residuals

3.5 Adjust costs

Because the concept has a regular scan resolution, nine printers that have a regular scan resolutions are chosen for constructing the distribution of the cost of concept. The costs of these nine printers are adjusted for the difference between their maximum copy speed and that of the concept parallel to the regression line, as illustrated in Fig. 9.

3.6 Construct histogram of adjusted cost and fit a theoretical distribution

Figure 10 shows the histogram of the adjusted costs and the normal distribution fitted to the histogram. The distribution of the costs adjusted parallel to the regression line is the same as the distribution of residuals obtained from the regression analysis (except that the mean of the adjusted costs is not zero and is estimated by the regression model). Because the assumption of the normally distributed residuals is supported in Step 4, the distribution of the adjusted costs can be approximated by a normal distribution. This distribution serves as the distribution of the cost of concept.

4 Comparison of adaptation approaches

In addition to illustrating the proposed CBR approach for constructing a distribution of the cost of concept, this section compares the accuracy of point estimates for three cost adjustment approaches: no-adjustment, linear-adjustment, and parallel-adjustment. No-adjustment is the most popular approach in the CBR applications for cost estimation (Shepperd and Schofield 1997; Mendes et al. 2003;

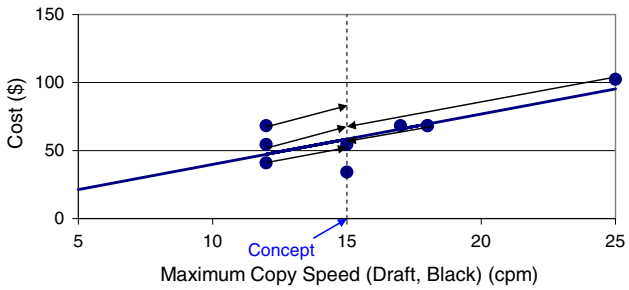

Fig. 9 Cost adjustment

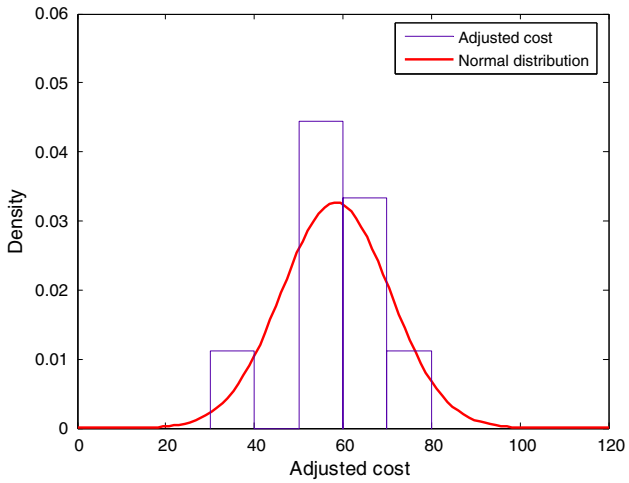

Fig. 10 Cost distribution

Kim et al. 2004; An et al. 2007). In this approach, point estimates (e.g., averages) are calculated from the cost of the retrieved products without adjusting them before the calculation. Linear-adjustment first identifies a feature (specification) of the retrieved products that has the largest correlation with the cost. Then the costs of the retrieved products are adjusted according to the ratio of this feature of the concept and that of the retrieved products. The point estimates (Jeffery et al. 2000) or cost distributions (Takai 2007) can be obtained from these adjusted costs. Parallel-adjustment corresponds to the approach proposed in this paper in which the retrieved costs are adjusted parallel to the regression line (or surface if there are more than one numeric cost relevant specifications). Figure 11 schematically illustrates these three cost adjustment approaches when there is only one numeric cost relevant feature.

The accuracy of point estimation is evaluated by applying leave-one-out cross validation to the retrieved ten printers: P1–P10. In the leave-one-out cross validation, one printer is hypothesized as the concept and is separated from the remaining nine printers. The cost of this hypothetical concept is estimated from the remaining nine printers and the estimated cost $\hat{c}$ is compared with the actual cost c. This process is repeated ten times, each time hypothesizing that one of the ten printers is the concept.

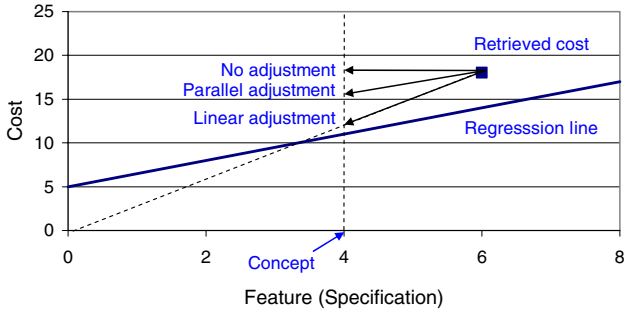

Fig. 11 Illustrations of the cost adjustment approaches

The accuracy of point estimation is evaluated by the "mean magnitude of relative error" (MMRE) in Eq. 3 (Shepperd and Schofield 1997; Mendes et al. 2003; Jeffery et al. 2000; Kim et al. 2004; An et al. 2007). MMRE is the average of the "magnitude of relative error" (MRE) $\left|\frac{c_j-\hat{c}_j}{c_j}\right|$ that measures the relative absolute difference of the actual cost c_j and the estimated cost $\hat{c}_j$ for each project ($j = 1, 2,\ldots, m$).

$$\text{MMRE} = \frac{1}{m}\sum_{j=1}^{m}\left|\frac{c_j - \hat{c}_j}{c_j}\right|. \tag{3}$$

Figure 12 summarizes the MMRE for three cost adjustment approaches. In the case of no-adjustment, point estimates are calculated for the closest one, two, three, four, and nine neighbors in order to optimize the number of retrieved printers. In this example, the closest three neighbors give the minimum MMRE for the no-adjustment approach. Comparing the three approaches, the parallel-adjustment gives the minimum MMRE 0.246 compared to that of the linear-adjustment, which is 0.257, and that of the no-adjustment (with three closest neighbors), which is 0.254.

5 Conclusions and future work

This paper proposed a CBR approach toward developing a belief (constructing a distribution) about the cost of concept (the cost of the final product developed from a concept) utilizing the data in the knowledge-base. The assumption in this paper is that a product concept is selected before deciding the detail design and manufacturing processes (Fig. 1). Primary functions, technologies, and performance specifications of a concept (e.g., printing speed) are specified in the concept selection stage; however, design specifications (e.g., weight, dimensions, product architecture, material used) are not decided in this stage.

The CBR approach consists of storing knowledge, defining a new problem, retrieving cases from the knowledge-base that are the most similar to the new problem, and adapting the solution of the retrieved cases to the new problem. In this paper, the new problem is the construction of a distribution and the estimation of the cost of concept, and cases are the products in the knowledge-base.

In this paper, three hierarchical clustering methods (complete linkage method, average linkage method and Ward's method) were used to retrieve from a heterogeneous knowledge-base a highly homogeneous set of products that are similar to the concept. The costs of the retrieved products are adjusted for differences in cost relevant specifications of the concept and those of the retrieved products using a regression model. The distribution of the cost of concept is constructed by fitting a normal distribution to the adjusted costs.

In addition to constructing distributions, this paper compared the accuracy of point estimates obtained from three cost adjustment approaches (no-adjustment, linear-adjustment, and parallel-adjustment) using leave-one-out cross validation. Parallel-adjustment corresponds to the approach proposed in this paper that uses regression model. Comparing three adaptation approaches, parallel-adjustment resulted in the smallest average estimation error; however, more research is needed to generalize this conclusion.

Thus the first future work is to compare different methodologies to conclude an optimum set of

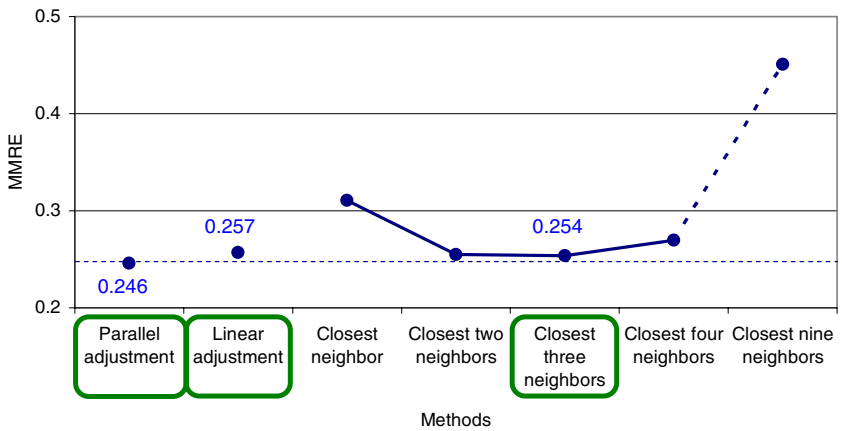

Fig. 12 Comparison of the accuracy of point estimates

methodologies in the CBR framework. For example, in addition to hierarchical clustering methods used in this paper, other classification methods such as *k*-means method classification methods (e.g., *k*-nearest neighbor, decision trees, and neural networks) (Hastie et al. 2001) may be used to retrieve similar systems. Similarly, in addition to fitting theoretical distribution, density estimation (Silverman 1986; Hastie et al. 2001) may be used for constructing distribution.

Another future work is to integrate technology forecasting in constructing cost distribution of a concept. When engineers consider using an innovative technology in a new product that has not been used for the products in the knowledge-base, simply using the knowledge-base may not be sufficient to accurately construct distributions and estimate costs of innovative concepts. Thus, forecasting the effects of new technologies needs to be integrated with estimating costs and constructing cost distributions of concepts.

Finally, the long-term objective of this research is to establish a methodology that can automatically construct a knowledge-base, retrieve similar cases from the knowledge-base, identify cost relevant specifications (parameters), construct cost distributions, and estimate costs, while using as minimum human judgment as possible. Integration of the proposed methodology into a decision support system is hence another future work.

Appendix A: Clusters of products similar to the concept

See Fig. 13.

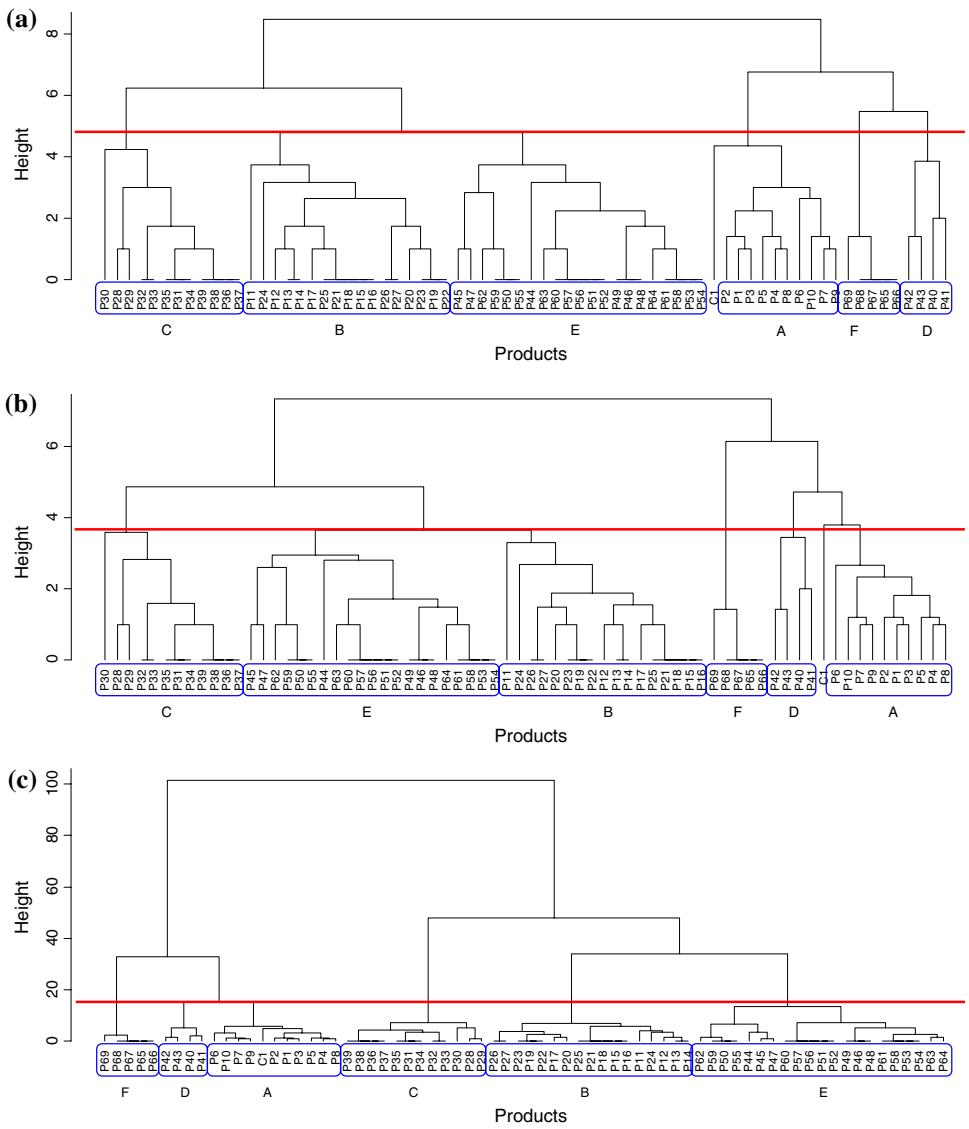

Fig. 13 Dendrogram of the concept and the products in the knowledge-base. **a** Complete linkage method. **b** Average linkage method. **c** Ward's method

References

Al-Shihabi T, Zeid I (1998) A design-plan-oriented methodology for applying case-based adaptation to engineering design. Artif Intell Eng Des Anal Manuf 12(5):463–478. doi:10.1017/S0890060498125052

An S-H, Kim G-H, Kang K-I (2007) A case-based reasoning cost estimating model using experience by analytic hierarchy process. Build Environ 42(7):2573–2579. doi:10.1016/j.buildenv.2006.06.007

Angelis L, Stamelos I (2000) A simulation tool for efficient analogy based cost estimation. Empir Softw Eng 5(1):35–68. doi:10.1023/A:1009897800559

Auer M, Trendowicz A, Graser B, Haunschmid E, Biffl S (2006) Optimal project feature weights in analogy-based cost estimation: improvement and limitations. IEEE Trans Softw Eng 32(2):83–92. doi:10.1109/TSE.2006.1599418

Bardasz T, Zeid I (1991) Applying analogical problem solving to mechanical design. Comput Aided Des 23(3):202–212. doi:10.1016/0010-4485(91)90090-J

Bardasz T, Zeid I (1993) DEJAVU: case-based reasoning for mechanical design. Artif Intell Eng Des Anal Manuf 7(2):111–124

Brimson JA (1998) Feature costing: beyond ABC. J Cost Manage (January/February):6–12

Clausing D (1993) Total quality development: a step-by-step guide to world-class concurrent engineering. ASME Press, New York

Cooper R, Kaplan RS (1987) How cost accounting systematically distorts product costs. In: Kaplan RS, Bruns WJ Jr (eds) Accounting and management: field study perspectives. Harvard Business School Press, Cambridge

Curran R, Price M, Raghunathan S, Benard E, Crosby S, Castagne S, Mawhinney P (2005) Integrating aircraft cost modeling into conceptual design. Concurrent Engineering. Res Appl 13(4):321–330

Dewhuest P, Boothroyd G (1988) Early cost estimating in product design. J Manuf Syst 7(3):183–191. doi:10.1016/0278-6125(88)90003-9

Hamaker J (1995) Parametric estimating. In: Stewart RD, Wyskida RM, Johannes JD (eds) Cost estimator's reference manual. Wiley, New York

Hastie T, Tibshirani R, Friedman J (2001) The elements of statistical learning: data mining, inference, and prediction. Springer, New York

Hauser J, Clausing D (1988) The house of quality. Harv Bus Rev 66(3):63–73

Hicks BJ, Culley SJ, Mullineux G (2002) Cost estimation for standard components and systems in the early phases of the design process. J Eng Des 13(4):271–292. doi:10.1080/0954482021000050802

Jeffery R, Ruhe M, Wieczorek I (2000) A comparative study of two software development cost modeling techniques using multi-organizational and company-specific data. Inf Softw Technol 42(14):1009–1016. doi:10.1016/S0950-5849(00)00153-1

Joint Industry/Government Parametric Cost Estimating Initiative (1999) Parametric estimating handbook. Department of Defense

Kim G-H, An S-H, Kang K-I (2004) Comparison of construction cost estimating models based on regression analysis, neural networks, and case-based reasoning. Build Environ 39(10):1235–1242. doi:10.1016/j.buildenv.2004.02.013

Kolodner JL (1993) Case-based reasoning. Morgan Kaufmann Publishers, San Mateo

Lee KH, Lee K-Y (2002) Agent-based collaborative design system and conflict resolution based on a case-based reasoning approach. Artif Intell Eng Des Anal Manuf 16(2):93–102. doi:10.1017/S0890060402020085

Looney SW, Gulledge TR Jr (1985) Use of the correlation coefficient with normal probability plots. Am Stat 39(1):75–79. doi:10.2307/2683917

Maher ML, Zhang DM (1993) CADSYN: a case-based design process model. Artif Intell Eng Des Anal Manuf 7(2):97–110

Mardia KV, Kent JT, Bibby JM (2000) Multivariate analysis, 7th edn. Academic Press, San Diego

Mauchand M, Siadat A, Bernard A, Perry N (2008) Proposal for tool-based method of product cost estimation during conceptual design. J Eng Des 19(2):159–172. doi:10.1080/09544820701802857

Mendes E, Watson I, Triggs C, Mosley N, Counsell S (2003) A comparative study of cost estimation models for web hypermedia applications. Empir Softw Eng 8(2):163–196. doi:10.1023/A:1023062629183

Mileham AR, Currie GC, Miles AW, Bradford DT (1993) A parametric approach to cost estimating at the conceptual stage of design. J Eng Des 4(2):117–125. doi:10.1080/09544829308914776

Neter J, Kutner MH, Nachtsheim CJ, Wasserman W (1996) Applied linear statistical models. Irwin, Chicago

Newnes LB, Mileham AR, Cheung WM, Marsh R, Lanham JD, Saravi ME, Bradbery RW (2008) Predicting the whole-life cost of a product at the conceptual design stage. J Eng Des 19(2):99–112. doi:10.1080/09544820701803061

Otto KN, Wood KL (2001) Product design: techniques in reverse engineering and new product development. Prentice-Hall Inc, Upper Saddle River

Pahl G, Beitz W (1996) Engineering design—a systematic approach, 2nd edn. Springer, London

Pugh S (1990) Total design. Addison-Wesley, Reading

Reich Y, Kapeliuk A (2004) Case-based reasoning with subjective influence knowledge. Appl Artif Intell 18(8):735–760. doi:10.1080/08839510490496978

Roderman S, Tsatsoulis C (1993) PANDA: a case-based system to AI novice designers. Artif Intell Eng Des Anal Manuf 7(2):125–133

Rosenman M (2000) Case-based evolutionary design. Artif Intell Eng Des Anal Manuf 14(1):17–29. doi:10.1017/S0890060400141022

Roy R, Kelvesjo S, Forsberg S, Rush C (2001) Quantitative and qualitative cost estimating for engineering design. J Eng Des 12(2):147–162. doi:10.1080/09544820110038997

Seo K-K, Park J-H, Jang D-S, Wallace D (2002) Approximate estimation of the product life cycle cost using artificial neural network in conceptual design. Int J Adv Manuf Technol 19(6):461–471. doi:10.1007/s001700200049

Shepperd M, Schofield C (1997) Estimating software project effort using analogies. IEEE Trans Softw Eng 23(12):736–743. doi:10.1109/32.637387

Shiva Kumar H, Krishnamoorthy CS (1995) A framework for case-based reasoning in engineering design. Artif Intell Eng Des Anal Manuf 9(3):161–182

Silverman BW (1986) Density estimation for statistics and data analysis. Chapman & Hall, New York

Takai S (2007) An approach toward estimating cost of a new system in the conceptual design stage using knowledge-base. In: Proc ASME DETC/CIE, Las Vegas, Nevada, USA, DETC2007-35072

Ulrich KT, Eppinger SD (2004) Product design and development. McGraw-Hill, Boston

Watson P, Curran R, Murphy A, Cowan S (2006) Cost estimation of machined parts within and aerospace supply chain. Concurrent Engineering. Res Appl 14(1):17–26

Wood WH, Agogino AM (1996) Case-based conceptual design information server for concurrent engineering. Comput Aided Des 28(5):361–369. doi:10.1016/0010-4485(95)00055-0

Wyskida RM (1995) Statistical techniques in cost estimation. In: Stewart RD, Wyskida RM, Johannes JD (eds) Cost estimator's reference manual. Wiley, New York