Int J Adv Manuf Technol (2006) 27: 774–780
DOI 10.1007/s00170-004-2250-0

ORIGINAL ARTICLE

Muh-Cherng Wu · Shih-Ching Wu · Tai-Chang Hsia · Shang-Hwa Hsu

A similarity inference method for reducing the cost of pair comparison

Received: 12 February 2003 / Accepted: 13 May 2004 / Published online: 23 February 2005

Abstract Group technology must group similar parts into families. In classifying parts based on their global shapes, the similarity of parts has to be manually measured by performing pair comparison. The cost of exhaustively performing pair comparison is quite high when the number of parts to be grouped is large. This paper proposes interval intersection, a novel similarity inference method that effectively infers the pair-comparison data from a set of known data. Justified by empirical experiments, the proposed method outperforms the previous methods when 31% or more of data is known.

Keywords Comparison · Group technology · Pair · Set intersection · Similarity inference

1 Introduction

Group technology (GT) is a technique for enhancing design/ manufacturing productivity by grouping similar workpieces into families [1]. Much GT research has been done on the automatic classification of workpieces, which use local features such as holes, steps, and slots as the criteria for justifying similar ity [2–5]. Yet, similarity in local features does not ensure similarity in global shape. These local feature-based GT systems are therefore limited in applications.

Recent GT research aims to develop automatic workpiece classification systems based on the similarity of global shapes [6–10]. These global-shape-based systems are particularly well-suited to the application of retrieving similar designs. Yet, automatically characterizing the global shapes of workpieces in a computational form needs an algorithm, which may not reflect users' recognition model. The effectiveness of such characterization should therefore be justified by user's judgments.

Hsu et al. proposed a full-data benchmarking method for justifying the effectiveness of a global-shape-based GT system [11]. A set of sample workpieces is selected from the general population of workpieces. Then, subjects exhaustively make pair comparisons of the sample workpieces based on their global shape similarity. The full pair-comparison data is used as a benchmark for measuring the effectiveness of an automatic workpiece classification system.

Yet, establishing the full pair comparison data is a time-consuming and labor-intensive task. Suppose the number of sample workpieces is n, and the number of exhaustive pair comparison is $n(n-1)/2$. That is, 100 sample workpieces would need about 5,000 pair comparisons. To reduce the number of pair comparison, some research aims to use partial experiment data to infer the full pair-comparison data. For example, of the 5000 pair-comparison data, some studies can use a part of them, say 2000, to infer the remaining 3000 data.

Some similarity inference methods have been developed in the literature [12, 13]. These methods include: (1) the Hamming distance method [13, 14]; (2) the max-min method [12, 15]; (3) the interval average method [12]; and (4) the weighing interval average method [13]. However, these four methods are not accurate enough in inferring the unknown data.

This paper presents a more accurate, similarity inference method, called the interval intersection method. Our experiments show that the interval intersection method outperforms the previous four similarity inference methods in most cases. Thus, the accuracy in inferring the similarity data has been improved.

The remainder of this paper is organized as follows: Section 2 formulates the similarity inference problem and explains how the full pair-comparison data is obtained; Section 3 presents the interval intersection method; Section 4 introduces the previous four inference methods; Section 5 defines three metrics for comparing the effectiveness of the inference methods, and gives their comparison results; and Section 6 offers concluding remarks.

M.-C. Wu (✉) · S.-C. Wu · S.-H. Hsu
Department of Industrial Engineering and Management,
National Chaio Tung University,
Hsin-Chu, Taiwan, R.O.C.
E-mail: mcwu@cc.nctu.edu.tw
Fax: +886-3-5720610

T.-C. Hsia
Department of Industrial Engineering and Management,
Chien Kuo Institute of Technology,
Chan-Hua, Taiwan, R.O.C.

2 Problem formulation and experiment data collection

The problem of similarity inference can be illustrated in the format of a matrix. In Eq. 1, a matrix $S = [s_{ij}]$ $(1 \le i, j \le n)$ represents the full pair-comparison data of n sample workpieces, where $s_{ij} \in [0, 1]$ denotes the similarity between two objects, i and j. The higher the value of s_{ij}, the more similar are the two workpieces. Notice $s_{ii} = 1$ and $s_{ij} = s_{ji}$ for $1 \le i, j \le n$; here, matrix S is symmetric and each of its diagonal elements is 1. Suppose a portion of the matrix elements (s_{ij} for $1 \le i \le m; i \le j \le n$) is known, the problem of interest is to develop methods to infer the other elements (s_{ij} for $m+1 \le i \le n; i \le j \le n$). Referring to Eq. 1, the known data is outlined by the trapezoid and the data to be inferred is outlined by the triangle. The inferred matrix is represented by $\hat{S} = [\hat{s}_{ij}]$, where $\hat{s}_{ij} = s_{ij}(1 \le i \le m; i \le j \le n)$ and $\hat{s}_{ij}(m+1 \le i \le n; i \le j \le n)$ is the inferred data.

$$S = \begin{bmatrix} 1 & s_{12} & \cdot & \cdot & \cdot & \cdot & s_{1n} \\ & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot \\ & & 1 & \cdot & \cdot & \cdot & s_{mn} \\ & & & 1 & \cdot & \cdot & \cdot \\ & & & & \cdot & \cdot & \cdot \\ & & & & & \cdot & \cdot \\ & & & & & & 1 \end{bmatrix} . \qquad (1)$$

The full similarity pair comparison data of n sample workpieces are collected by the procedure proposed by Hsu et al. [11], which is explained by the example with 36 sample workpieces shown in Fig. 1. The procedure asks 30 subjects (users of the GT system in an aircraft manufacturing company) to make an exhaustive pair comparison. Each subject then has to do 630 (i.e. $36 \times 35/2$) pair comparisons. Each pair comparison is represented in one of the five fuzzy linguistic terms as follows: very high similarity, high similarity, medium similarity, low similarity, very low similarity [16]. Each linguistic term is modeled by a fuzzy number. For each pair comparison, a fuzzy operation is applied to aggregate the 30 fuzzy numbers into one fuzzy number [17]. The aggregated fuzzy number is subsequently converted to a crisp value by a defuzzification process [17]. The 36×36 matrix S for the sample workpieces can thus be obtained as shown in Table 1.

3 Proposed similarity inference methods

The proposed similarity inference method, the interval intersection method, is presented in this section. The problem of interest is described as follows: suppose s_{pq} is not known, while s_{pi} and s_{qi} are known for some i. Herein, the set including such i is denoted by T. The proposed similarity inference method is to determine $\hat{s}_{pq}$.

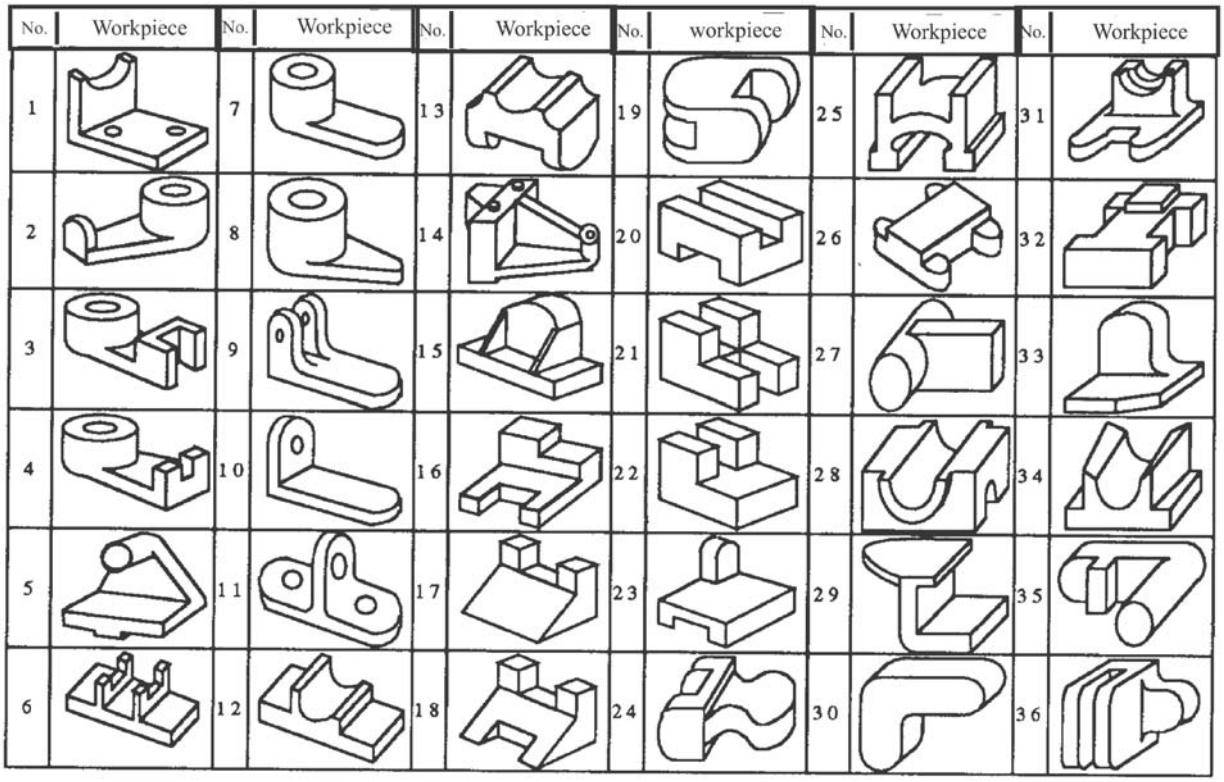

Fig. 1. The 36 sample workpieces used for pair comparison

Table 1. Complete experimental data

	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21	22	23	24	25	26	27	28	29	30	31	32	33	34	35	36
1	1.000	0.403	0.356	0.467	0.250	0.464	0.428	0.300	0.489	0.650	0.558	0.425	0.236	0.294	0.275	0.389	0.339	0.369	0.165	0.272	0.325	0.397	0.325	0.221	0.286	0.250	0.289	0.327	0.289	0.380	0.536	0.311	0.517	0.419	0.249	0.225
2	0.403	1.000	0.742	0.789	0.389	0.225	0.792	0.762	0.508	0.581	0.428	0.241	0.316	0.383	0.306	0.188	0.136	0.145	0.350	0.162	0.162	0.216	0.322	0.406	0.204	0.303	0.542	0.208	0.311	0.280	0.333	0.196	0.444	0.227	0.542	0.356
3	0.356	0.742	1.000	0.823	0.408	0.286	0.675	0.667	0.397	0.411	0.403	0.250	0.252	0.361	0.328	0.336	0.302	0.241	0.369	0.381	0.358	0.414	0.314	0.367	0.202	0.222	0.517	0.221	0.258	0.210	0.389	0.317	0.331	0.230	0.456	0.361
4	0.467	0.789	0.823	1.000	0.494	0.336	0.759	0.725	0.483	0.483	0.428	0.250	0.247	0.417	0.300	0.408	0.344	0.303	0.272	0.303	0.316	0.406	0.283	0.361	0.213	0.244	0.572	0.199	0.261	0.207	0.389	0.283	0.311	0.227	0.567	0.308
5	0.250	0.389	0.408	0.494	1.000	0.280	0.397	0.397	0.225	0.250	0.230	0.286	0.250	0.667	0.403	0.275	0.289	0.294	0.199	0.241	0.252	0.258	0.317	0.250	0.213	0.241	0.461	0.225	0.436	0.244	0.297	0.210	0.342	0.208	0.558	0.322
6	0.464	0.225	0.286	0.336	0.280	1.000	0.179	0.162	0.350	0.314	0.322	0.642	0.378	0.185	0.291	0.597	0.425	0.433	0.202	0.561	0.575	0.608	0.369	0.224	0.381	0.428	0.185	0.366	0.238	0.151	0.347	0.564	0.264	0.567	0.230	0.272
7	0.428	0.792	0.675	0.759	0.397	0.179	1.000	0.818	0.622	0.667	0.533	0.239	0.322	0.380	0.308	0.222	0.131	0.142	0.336	0.142	0.170	0.204	0.261	0.392	0.170	0.283	0.642	0.191	0.241	0.378	0.353	0.168	0.469	0.196	0.606	0.347
8	0.300	0.762	0.667	0.725	0.397	0.162	0.818	1.000	0.508	0.539	0.442	0.191	0.247	0.389	0.331	0.156	0.151	0.145	0.297	0.107	0.153	0.153	0.250	0.344	0.133	0.227	0.575	0.185	0.367	0.347	0.342	0.136	0.381	0.185	0.575	0.347
9	0.489	0.508	0.397	0.483	0.225	0.350	0.622	0.508	1.000	0.798	0.736	0.330	0.394	0.264	0.339	0.241	0.215	0.221	0.467	0.216	0.249	0.333	0.331	0.556	0.182	0.328	0.367	0.168	0.264	0.364	0.461	0.207	0.475	0.227	0.428	0.444
10	0.650	0.581	0.411	0.483	0.250	0.314	0.667	0.539	0.798	1.000	0.781	0.291	0.314	0.255	0.397	0.260	0.172	0.184	0.408	0.185	0.227	0.269	0.400	0.444	0.193	0.330	0.436	0.167	0.319	0.558	0.483	0.199	0.594	0.219	0.494	0.433
11	0.558	0.428	0.403	0.428	0.230	0.322	0.533	0.442	0.736	0.781	1.000	0.319	0.286	0.255	0.325	0.156	0.095	0.147	0.389	0.151	0.142	0.173	0.331	0.436	0.170	0.381	0.431	0.207	0.286	0.450	0.417	0.151	0.536	0.272	0.442	0.394
12	0.425	0.241	0.250	0.250	0.286	0.642	0.239	0.191	0.330	0.291	0.319	1.000	0.569	0.199	0.411	0.342	0.330	0.339	0.286	0.375	0.333	0.392	0.367	0.327	0.531	0.353	0.225	0.575	0.417	0.224	0.403	0.375	0.314	0.864	0.269	0.319
13	0.236	0.316	0.252	0.247	0.250	0.378	0.322	0.247	0.394	0.314	0.286	0.569	1.000	0.190	0.447	0.266	0.238	0.249	0.684	0.405	0.303	0.328	0.386	0.628	0.422	0.275	0.269	0.453	0.358	0.286	0.411	0.227	0.406	0.531	0.258	0.500
14	0.294	0.383	0.361	0.417	0.667	0.185	0.380	0.389	0.264	0.255	0.255	0.199	0.190	1.000	0.408	0.247	0.257	0.280	0.182	0.181	0.269	0.289	0.286	0.239	0.174	0.176	0.353	0.199	0.252	0.224	0.339	0.264	0.289	0.261	0.369	0.283
15	0.275	0.306	0.328	0.300	0.403	0.291	0.308	0.331	0.339	0.397	0.325	0.411	0.447	0.408	1.000	0.367	0.422	0.444	0.378	0.272	0.291	0.305	0.558	0.453	0.205	0.367	0.336	0.272	0.367	0.311	0.517	0.277	0.681	0.353	0.319	0.464
16	0.389	0.188	0.336	0.408	0.275	0.597	0.222	0.156	0.241	0.260	0.156	0.342	0.266	0.247	0.367	1.000	0.589	0.639	0.300	0.600	0.742	0.753	0.469	0.247	0.303	0.258	0.199	0.258	0.297	0.244	0.531	0.569	0.408	0.300	0.275	0.347
17	0.339	0.136	0.302	0.344	0.289	0.425	0.131	0.151	0.215	0.172	0.095	0.330	0.238	0.257	0.422	0.589	1.000	0.864	0.255	0.500	0.569	0.633	0.350	0.227	0.275	0.241	0.188	0.236	0.308	0.247	0.394	0.464	0.305	0.286	0.239	0.258
18	0.369	0.145	0.241	0.303	0.294	0.433	0.142	0.145	0.221	0.184	0.147	0.339	0.249	0.280	0.444	0.639	0.864	1.000	0.289	0.525	0.653	0.661	0.383	0.252	0.325	0.252	0.176	0.255	0.303	0.207	0.481	0.489	0.283	0.369	0.196	0.250
19	0.165	0.350	0.369	0.272	0.199	0.202	0.336	0.297	0.467	0.408	0.389	0.286	0.684	0.182	0.378	0.300	0.255	0.289	1.000	0.506	0.378	0.361	0.353	0.717	0.341	0.366	0.266	0.367	0.328	0.294	0.381	0.213	0.353	0.289	0.247	0.639
20	0.272	0.162	0.381	0.303	0.241	0.561	0.142	0.107	0.216	0.185	0.151	0.375	0.405	0.181	0.272	0.600	0.500	0.525	0.506	1.000	0.700	0.700	0.450	0.328	0.406	0.319	0.188	0.456	0.356	0.151	0.311	0.533	0.225	0.369	0.179	0.394
21	0.325	0.162	0.358	0.316	0.252	0.575	0.170	0.153	0.249	0.227	0.142	0.333	0.303	0.269	0.291	0.742	0.569	0.653	0.378	0.700	1.000	0.807	0.444	0.275	0.333	0.269	0.202	0.277	0.300	0.297	0.458	0.528	0.339	0.322	0.266	0.339
22	0.397	0.216	0.414	0.406	0.258	0.608	0.204	0.153	0.333	0.269	0.173	0.392	0.328	0.289	0.305	0.753	0.633	0.661	0.361	0.700	0.807	1.000	0.411	0.336	0.339	0.306	0.244	0.278	0.305	0.269	0.414	0.572	0.336	0.375	0.247	0.394
23	0.325	0.322	0.314	0.283	0.317	0.369	0.261	0.250	0.331	0.400	0.331	0.367	0.386	0.286	0.558	0.469	0.350	0.383	0.353	0.450	0.444	0.411	1.000	0.392	0.364	0.461	0.286	0.342	0.478	0.297	0.486	0.464	0.625	0.364	0.325	0.467
24	0.221	0.406	0.367	0.361	0.250	0.224	0.392	0.344	0.556	0.444	0.436	0.327	0.628	0.239	0.453	0.247	0.227	0.252	0.717	0.328	0.275	0.336	0.392	1.000	0.294	0.347	0.422	0.300	0.336	0.352	0.461	0.222	0.492	0.297	0.291	0.745
25	0.286	0.204	0.202	0.213	0.213	0.381	0.170	0.133	0.182	0.193	0.170	0.531	0.422	0.174	0.205	0.303	0.275	0.325	0.341	0.406	0.333	0.339	0.364	0.294	1.000	0.241	0.176	0.531	0.406	0.170	0.342	0.322	0.241	0.567	0.196	0.311
26	0.250	0.303	0.222	0.244	0.241	0.428	0.283	0.227	0.328	0.330	0.381	0.353	0.275	0.176	0.367	0.258	0.241	0.252	0.366	0.319	0.269	0.306	0.461	0.347	0.241	1.000	0.367	0.236	0.209	0.300	0.392	0.597	0.428	0.325	0.322	0.325
27	0.289	0.542	0.517	0.572	0.461	0.185	0.642	0.575	0.367	0.436	0.431	0.225	0.269	0.353	0.336	0.199	0.188	0.176	0.266	0.188	0.202	0.244	0.286	0.422	0.176	0.367	1.000	0.188	0.277	0.342	0.325	0.181	0.369	0.207	0.664	0.333
28	0.327	0.208	0.221	0.199	0.225	0.366	0.191	0.185	0.168	0.167	0.207	0.575	0.453	0.199	0.272	0.258	0.236	0.255	0.367	0.456	0.277	0.278	0.342	0.300	0.531	0.236	0.188	1.000	0.280	0.176	0.358	0.224	0.216	0.656	0.193	0.292
29	0.289	0.311	0.258	0.261	0.436	0.238	0.241	0.367	0.264	0.319	0.286	0.417	0.358	0.252	0.367	0.297	0.308	0.303	0.328	0.356	0.300	0.305	0.478	0.336	0.406	0.209	0.277	0.280	1.000	0.389	0.381	0.272	0.533	0.375	0.255	0.467
30	0.380	0.280	0.210	0.207	0.244	0.151	0.378	0.347	0.364	0.558	0.450	0.224	0.286	0.224	0.311	0.244	0.247	0.207	0.294	0.151	0.297	0.269	0.297	0.352	0.170	0.300	0.342	0.176	0.389	1.000	0.389	0.145	0.478	0.188	0.439	0.353
31	0.536	0.333	0.389	0.389	0.297	0.347	0.353	0.342	0.461	0.483	0.417	0.403	0.411	0.339	0.517	0.531	0.394	0.481	0.381	0.311	0.458	0.414	0.486	0.461	0.342	0.392	0.325	0.358	0.381	0.389	1.000	0.280	0.622	0.425	0.333	0.447
32	0.311	0.196	0.317	0.283	0.210	0.564	0.168	0.136	0.207	0.199	0.151	0.375	0.227	0.264	0.277	0.569	0.464	0.489	0.213	0.533	0.528	0.572	0.464	0.222	0.322	0.597	0.181	0.224	0.272	0.145	0.280	1.000	0.316	0.311	0.233	0.219
33	0.517	0.444	0.331	0.311	0.342	0.264	0.469	0.381	0.475	0.594	0.536	0.314	0.406	0.289	0.681	0.408	0.305	0.283	0.353	0.225	0.339	0.336	0.625	0.492	0.241	0.428	0.369	0.216	0.533	0.478	0.622	0.316	1.000	0.344	0.361	0.561
34	0.419	0.227	0.230	0.227	0.208	0.567	0.196	0.185	0.227	0.219	0.272	0.864	0.531	0.261	0.353	0.300	0.286	0.369	0.289	0.369	0.322	0.375	0.364	0.297	0.567	0.325	0.207	0.656	0.375	0.188	0.425	0.311	0.344	1.000	0.213	0.277
35	0.249	0.542	0.456	0.567	0.558	0.230	0.606	0.575	0.428	0.494	0.442	0.269	0.258	0.369	0.319	0.275	0.239	0.196	0.247	0.179	0.266	0.247	0.325	0.291	0.196	0.322	0.664	0.193	0.255	0.439	0.333	0.233	0.361	0.213	1.000	0.339
36	0.225	0.356	0.361	0.308	0.322	0.272	0.347	0.347	0.444	0.433	0.394	0.319	0.500	0.283	0.464	0.347	0.258	0.250	0.639	0.394	0.339	0.394	0.467	0.745	0.311	0.325	0.333	0.292	0.467	0.353	0.447	0.219	0.561	0.277	0.339	1.000

3.1 The goal of the algorithm

The goal of the proposed inference method is to estimate s_{pq} through another workpiece $k \in T$. Let $\hat{s}_{pq(k)}$ denote the estimation of s_{pq} through workpiece k. The upper bound and the lower bound of $\hat{s}_{pq(k)}$ are respectively represented by $\hat{s}^U_{pq(k)}$ and $\hat{s}^L_{pq(k)}$. The proposed inference method involves a two-step sequence: (1) for each object $k \in T$, the bounding interval $[\hat{s}^L_{pq(k)}, \hat{s}^U_{pq(k)}]$ is first computed (Fig. 2); and (2) $\hat{s}_{pq}$ is then determined by aggregating the bounding interval $[\hat{s}^L_{pq(k)}, \hat{s}^U_{pq(k)}]$ for all $k \in T$ (Fig. 3).

The goal of estimating $[\hat{s}^L_{pq(k)}, \hat{s}^U_{pq(k)}]$ is modeling the information contents of the global shape of a workpiece by a set. For a workpiece p, the information content of its global shape is represented by the set X_p, where the area of the set is $Area(X_p) = 1$. The similarity between two workpieces p and q is the intersection of two sets, that is, $s_{pq} = Area(X_p \cap X_q)$.

3.2 Deriving the bounding interval of $\hat{s}_{pq(k)}$

The proposed inference method first computes the bounding interval of $\hat{s}_{pq(k)}$; that is, $[\hat{s}^L_{pq(k)}, \hat{s}^U_{pq(k)}]$.

The lower bound of $\hat{s}_{pq(k)}$ would appear if $(X_p \cap X_q)$ is made as small as possible. We thus can infer that $(X_p \cap X_q)$ should be a subset of its relevant sets; that is, $(X_p \cap X_q) \subset (X_p \cap X_k)$ and $(X_p \cap X_q) \subset (X_q \cap X_k)$. This further implies that $\hat{s}^L_{pq(k)} = s_{pk} + s_{kq} - 1$ for $X_p \cap X_q \neq \Phi$, as shown in Fig. 4a, and $\hat{s}^L_{pq(k)} = 0$ for $X_p \cap X_q = \Phi$, as shown in Fig. 4b. The case in Fig. 4a implies that $s_{pk} + s_{kq} \geq 1$, and that in Fig. 4b implies that $s_{pk} + s_{kq} < 1$. By integrating the two cases, $\hat{s}^L_{pq(k)}$ can be computed as follows.

$$\hat{s}^L_{pq(k)} = Max(s_{pk} + s_{kq} - 1, 0)\,. \tag{2}$$

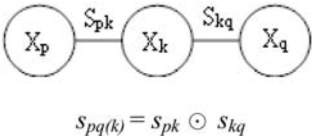

Fig. 2. Inference of s_{pq} through s_{pk} and s_{kq}, where symbol $\odot$ represents a similarity inference method

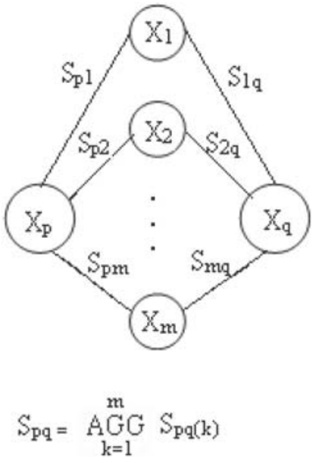

Fig. 3. s_{pq} is inferred by aggregating multiple $s_{pq(k)}$, where AGG represents a method of aggregation

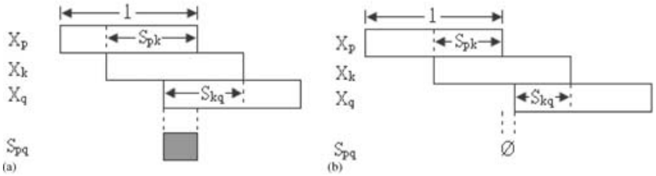

Fig. 4. Modeling lower bound of $\hat{s}_{pq(k)}$

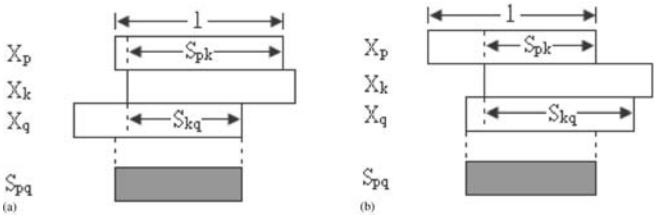

Fig. 5. Modeling upper bound of $\hat{s}_{pq(k)}$

The upper bound of $\hat{s}_{pq(k)}$ would appear if $(X_p \cap X_q)$ is made as large as possible. We can thus infer that$(X_p \cap X_q)$ is not a subset of its two relevant sets; that is, $(X_p \cap X_q) \not\subset (X_p \cap X_k)$, and $(X_p \cap X_q) \not\subset (X_q \cap X_k)$. This further implies that $\hat{s}^U_{pq(k)} = s_{qk} + (1 - s_{pq})$ when $s_{pk} \geq s_{kq}$, as shown in Fig. 5a, and $\hat{s}^U_{pq(k)} = s_{pk} + (1 - s_{kq})$, when $s_{pq} < s_{qk}$ as shown in Fig. 5b. By integrating the two cases in Fig. 5, $\hat{s}^U_{pq(k)}$can be computed as follows:

$$\hat{s}^U_{pq(k)} = 1 - |s_{pk} - s_{kq}|\,. \tag{3}$$

3.3 Aggregation of bounding intervals

For each object $k \in T$, there exists a bounding interval $[\hat{s}^L_{pq(k)}, \hat{s}^U_{pq(k)}]$ for $\hat{s}_{pq(k)}$. Let $[\hat{s}^L_{pq}, \hat{s}^U_{pq}]$ denote the bounding interval of $\hat{s}_{pq}$, which can be computed by the following aggregation procedure (where the individual bounding intervals are aggregated by the set intersection operation):

$$[\hat{s}^L_{pq}, \hat{s}^U_{pq}] = \bigcap_{k\in T} [\hat{s}^L_{pq(k)}, \hat{s}^U_{pq(k)}] = [\underset{k\in T}{Max}\, \hat{s}^L_{pq(k)}, \underset{k\in T}{Min}\, \hat{s}^U_{pq(k)}]\,.$$

$\hat{s}_{pq}$ is subsequently estimated as follows:

$$\begin{array}{ll} \hat{s}_{pq} = \dfrac{\hat{s}^U_{pq} + \hat{s}^L_{pq}}{2} & \text{if } \hat{s}^L_{pq} \leq \hat{s}^U_{pq} \text{ (i.e., } [\hat{s}^L_{pq}, \hat{s}^U_{pq}] \neq \Phi) \\ \hat{s}_{pq} = 0 & \text{if } \hat{s}^L_{pq} > \hat{s}^U_{pq} \text{ (i.e., } [\hat{s}^L_{pq}, \hat{s}^U_{pq}] = \Phi)\,. \end{array}$$

3.4 Example

The following example is used to illustrate the inference method. The example includes five sample workpieces: the first two rows ($m = 2$) of the S matrix, the known data, is denoted by S_2; and the last three rows of the S matrix, the unknown data, are to be inferred.

$$S_2 = \begin{bmatrix} 1 & 0.4 & 0.3 & 0.4 & 0.2 \\ & 1 & 0.7 & 0.7 & 0.3 \end{bmatrix}.$$

According to the above inference method, $\hat{s}^L_{34(1)} = \text{Max}\,(0.3+0.4-1, 0) = 0$ and $\hat{s}^U_{34(1)} = 1-|0.3-0.4| = 0.9$. Therefore, $[\hat{s}^L_{34(1)}, \hat{s}^U_{34(1)}] = [0, 0.9]$. Likewise, $[\hat{s}^L_{34(2)}, \hat{s}^U_{34(2)}] = [\text{Max}\,(0.7+0.7-1, 0), 1-|0.7-0.7|] = [0.4, 1]$. Then, $[\hat{s}^L_{34}, \hat{s}^U_{34}] = [0.4, 0.9]$ and $\hat{s}_{34} = \frac{0.4+0.9}{2} = 0.65$.

Accordingly, the symmetric matrix $\hat{S}$ can be estimated by the interval intersection method as follows:

$$\hat{S} = \begin{bmatrix} 1 & 0.4 & 0.3 & 0.4 & 0.2 \\ & 1 & 0.7 & 0.7 & 0.3 \\ & & 1 & 0.65 & 0.3 \\ & & & 1 & 0.3 \\ & & & & 1 \end{bmatrix}.$$

4 Previous inference methods

4.1 The Hamming distance method

The formulas for estimating $\hat{s}_{pq}$ by the Hamming distance method [13, 14] are described below:

$$\hat{s}_{pq(k)} = 1-|s_{pk}-s_{qk}|$$

$$\hat{s}_{pq} = \frac{\sum\limits_{k\in T} \hat{s}_{pq(k)}}{h}, \quad \text{where } h \text{ denotes the total number of objects in set } T.$$

Taking the example in Sect. 3.4, we can infer that $\hat{s}_{34(1)} = 1-|0.3-0.4| = 0.9$, $\hat{s}_{34(2)} = 1-|0.7-0.7| = 1.0$, and $\hat{s}_{34} = (1.0+0.9)/2 = 0.95$. By the Hamming distance method, matrix S is computed as follows:

$$\hat{S} = \begin{bmatrix} 1 & 0.4 & 0.3 & 0.4 & 0.2 \\ & 1 & 0.7 & 0.7 & 0.3 \\ & & 1 & 0.95 & 0.75 \\ & & & 1 & 0.7 \\ & & & & 1 \end{bmatrix}.$$

4.2 The max-min inference method

The max-min similarity inference method [12, 15] is introduced below:

$$\hat{s}_{pq(k)} = \text{Min}\,(s_{pk}, s_{kq})$$

$$\hat{s}_{pq} = \underset{k\in T}{\text{Max}}\,(\hat{s}_{pq(k)})\,.$$

Taking the example in Sect. 3.4, we can infer that $\hat{s}_{34(1)} = \min\,(0.3, 0.4) = 0.3$, $\hat{s}_{34(2)} = \min\,(0.7, 0.7) = 0.7$, and $\hat{s}_{34} = \max\,(0.3, 0.7) = 0.7$. By the max-min method, matrix S is computed as follows:

$$\hat{S} = \begin{bmatrix} 1 & 0.4 & 0.3 & 0.4 & 0.2 \\ & 1 & 0.7 & 0.7 & 0.3 \\ & & 1 & 0.7 & 0.3 \\ & & & 1 & 0.3 \\ & & & & 1 \end{bmatrix}.$$

4.3 The interval average method

The interval average method [13] is introduced below:

$$\hat{s}_{pq(k)} = \frac{\hat{s}^L_{pq(k)} + \hat{s}^U_{pq(k)}}{2}$$

$$\hat{s}_{pq} = \frac{\sum\limits_{k\in T} \hat{s}_{pq(k)}}{h}\,.$$

where h is the total number of objects in set T.

Taking the example in Sect. 3.4, we can infer that $\hat{s}_{34(1)} = (0+0.9)/2 = 0.45$, $\hat{s}_{34(2)} = (0.4+1) = 0.7$, and $\hat{s}_{34} = (0.45+0.7)/2 = 0.575$. By the interval average method, the matrix S is computed as follows:

$$\hat{S} = \begin{bmatrix} 1 & 0.4 & 0.3 & 0.4 & 0.2 \\ & 1 & 0.7 & 0.7 & 0.3 \\ & & 1 & 0.575 & 0.375 \\ & & & 1 & 0.35 \\ & & & & 1 \end{bmatrix}.$$

4.4 The weighting interval average method

The weighting interval average method [13] is introduced below:

$$\hat{s}_{pq(k)} = \frac{\hat{s}^L_{pq(k)} + \hat{s}^U_{pq(k)}}{2}$$

$$\hat{s}_{pq} = \frac{\sum\limits_{k\in T} \hat{s}_{pq(k)} \times [1-(\hat{s}^U_{pq(k)} - \hat{s}^L_{pq(k)})]}{\sum\limits_{k\in T} [1-(\hat{s}^U_{pq(k)} - \hat{s}^L_{pq(k)})]}\,.$$

Taking the example in Sect. 3.4, we can infer that $\hat{s}_{34(1)} = (0+0.9)/2 = 0.45$, $\hat{s}_{34(2)} = (0.4+1) = 0.7$, and $\hat{s}_{34} = \frac{0.45\times 0.1+0.7\times 0.4}{0.1+0.4} = 0.65$. By the weighted interval average method, the matrix S is computed as follows:

$$\hat{S} = \begin{bmatrix} 1 & 0.4 & 0.3 & 0.4 & 0.2 \\ & 1 & 0.7 & 0.7 & 0.3 \\ & & 1 & 0.65 & 0.33 \\ & & & 1 & 0.33 \\ & & & & 1 \end{bmatrix}.$$

5 Metrics and comparison

As mentioned above, the data to be inferred in matrix S involves s_{ij} for $m+1 \leq i \leq n; i \leq j \leq n$. Define $R = \{(i, j) \mid m+1 \leq i \leq n; i \leq j \leq n\}$. Then, s_{pq} is an element to be inferred $\forall (p, q) \in R$. For these matrix elements to be inferred, we would prefer that $\hat{s}_{ij}$ is as close to s_{ij} as possible. The following sections define three metrics to compare the effectiveness or accuracy of the above five inference methods.

5.1 Metrics

(A) Mean absolute deviation

The first metric, mean absolute deviation (MAD), is defined below:

$$\mathrm{MAD} = \frac{\sum\limits_{(p,q)\in R} |s_{pq} - \hat{s}_{pq}|}{r} \tag{4}$$

where r denotes the total number of elements in set R.

(B) Root mean square

The second metric, root mean square (RMS), is defined below:

$$\mathrm{RMS} = \sqrt{\frac{1}{r} \sum_{(p,q)\in R} (spq - \hat{s}pq)^2} \tag{5}$$

where r denotes the total number of elements in set R.

(C) Percentage of small deviation range

Let A and B be two sets defined as follows:

$$A = \{\hat{s}_{pq} \mid |s_{pq} - \hat{s}_{pq}| \leq 0.1\}$$
$$B = \{\hat{s}_{pq} \mid |s_{pq} - \hat{s}_{pq}| > 0.1\}$$

The third metric, PSD (percentage of small deviation) indicates the percentage of "good" inference, and is defined as follows:

$$\mathrm{PSD} = \frac{N(A)}{N(A) + N(B)} \tag{6}$$

where $N(A)$ denotes the number of elements in set A, $N(B)$ denotes the number of elements in set B.

Of the three metrics, the MAD and RMS measure the degree of accuracy for a similarity inference method. The smaller the two metrics, the higher the accuracy of estimation. The metric PSD measures the percentage of "good inference". The higher the PSD, the higher is the accuracy of estimation.

5.2 Comparison

The 36 workpieces shown in Fig. 1 are taken to compare the effectiveness of the five similarity inference methods. The full pair comparison data, the 36 × 36 symmetric matrix S, is shown in Table 1. As stated, a portion of the matrix elements s_{ij} for $1 \leq i \leq m$; $i \leq j \leq n$ is known, and the other elements s_{ij} for $m + 1 \leq i \leq n$; $i \leq j \leq n$ is to be inferred. Table 2 shows the percentage of known data for different m values.

Table 2. Percentage of known data in similarity inference

Value of m	$m = 36$	$m = 28$	$m = 24$	$m = 16$	$m = 12$	$m = 6$	$m = 1$
% of known data	100%	95%	90%	70%	56%	31%	6%

Table 3. MAD of the five inference methods

% of known data	Interval intersection method	Hamming distance method	Max-min method	Interval average method	Weighting interval average method
95%	0.052	0.505	0.116	0.105	0.084
90%	0.061	0.530	0.111	0.130	0.105
70%	0.056	0.505	0.099	0.119	0.092
56%	0.075	0.510	0.095	0.122	0.103
31%	0.089	0.504	0.102	0.126	0.107
6%	0.149	0.525	0.122	0.149	0.149

Table 4. RMS of the five inference methods

% of known data	Interval intersection method	Hamming distance method	Max-min method	Interval average method	Weighting interval average method
95%	0.069	0.513	0.137	0.125	0.100
90%	0.081	0.541	0.129	0.149	0.122
70%	0.073	0.516	0.117	0.138	0.110
56%	0.098	0.521	0.116	0.141	0.122
31%	0.111	0.519	0.127	0.147	0.127
6%	0.173	0.553	0.161	0.173	0.173

Table 5. PSD of the five inference methods

% of known data	Interval intersection method	Hamming distance method	Max-min method	Interval average method	Weighting interval average method
95%	82.1%	0%	53.6%	60.7%	60.7%
90%	78.8%	0%	45.5%	42.4%	48.48%
70%	86.3%	0%	53.2%	43.7%	57.9%
56%	74.6%	0%	55.8%	39.1%	50.7%
31%	62.1%	0.2%	54.9%	38.9%	50.1%
6%	32.6%	0.7%	49.6%	32.6%	32.6%

Table 3 depicts the MAD, Table 4 shows the RMS, and Table 5 describes the PSD for the five similarity inference methods. The proposed interval intersection method outperforms the other four methods in the three metrics when 31% or more data is known. Notice that when only 6% data is known, the proposed method is inferior to the max-min method. Yet, in such a scenario (where only 6% data known), the inferred results are not so "good". Table 5 shows that only 49.6% of the data inferred by the max-min method is "good". The table also shows that when 56% of data is known, 74.6% of inferred data (by the interval intersection method) is "good".

To ensure a reasonably accurate inference about pair comparison, according to the experiment, the known data had better be more than 56% or more. The proposed interval intersection method is the best choice for such an inference.

6 Conclusions

Building the pair comparison similarity matrix for a group of objects is helpful to some applications such as group technology. However, manually performing pair comparisons is quite time-consuming and labor-intensive.

This paper proposes interval intersection, a similarity inference method that can be used to infer the unknown data from a set of known data. According to three metrics, MAD, RMS and PSD, the proposed method is compared with the previous four methods in literature. Experiments show that the proposed method outperforms the other four when 31% or more data is known. The proposed method is inferior to the max-min method when only 6% data is known. Yet, when only 6% data is known, only about 50% of inferred data is "good". To ensure a reasonably accurate inference, the proposed interval intersection method is the best choice to date.

Acknowledgement The authors are grateful to the employees of the Aircraft Manufacturing, Aerospace Industrial Development Corporation (AIDC), Taichung, Taiwan, ROC, for their participation in the workpiece pair-comparison experiment.

References

1. Groover MP (2001) Automation, production systems, and computer integrated manufacturing, 2nd edn. Prentice-Hall, New York
2. Bhadra A, Fischer GW (1988) A new GT classification approach: a database with graphical dimensions. Manuf Rev 11:44–49
3. Henderson MR, Musti S (1988) Automated group technology part coding from a three-dimensional CAD database. Trans ASME J Eng Ind 110:278–287
4. Chen CS (1989) A form feature oriented coding scheme. Comput Ind Eng 11:227–233
5. Kaparthi S, Suresh N (1991) A neural network system for shape-based classification and coding of rotational parts. Int J Prod Res 29: 1771–1784
6. Lenau T, Mu L (1993) Features in integrated modeling of products and their production. Int J Comput Integ Manuf 6:65–73
7. Wu MC, Chen JR, Jen SR (1994) Global shape information modeling and classification of 2D workpieces. Int J Comput Integ Manuf 11: 325–335
8. Wu MC, Jen SR (1996) A neural network approach to the classification of 3D prismatic parts. Int J Adv Manuf Technol 11:325–335
9. Cheok BT, Zhang YF, Leow LF (1997) A skeleton-retrieving approach for the recognition of punched parts. Comput Ind 32:249–259
10. Gong BZ (2002) The processing of parts with group technology in an individual CNC machining center. J Mater Process Technol 129: 645–648
11. Hsu SH, Hsia TC, Wu MC (1997) A flexible classification method for evaluating the utility of automated workpiece classification system. Int J Adv Manuf Technol 13:637–648
12. Hsu SH, Hsia TC, Wu MC (1998) An efficient method for creating benchmark classification for automatic workpiece classification systems. Int J Adv Manuf Technol 14:481–494
13. Hsu SH, Hsia TC, Huang WY, Wu MC (2000) Enhancing the effectiveness of lean clustering in establishing benchmarks for automatic classification systems. Int J Manuf Technol Manage 1:288–317
14. Tan SK, Teh HH, Wang PZ (1994) Sequential representation of fuzzy similarity relations. Fuzzy Sets Sys 67:181–189
15. Zadeh LA (1973) Outline of a new approach to the analysis of complex systems and decision processes. IEEE Trans Sys Man Cybernetics 67:181–189
16. Chen SJ, Hwang CL (1992) Fuzzy multiple attribute decision making—methods and applications: a state-of-art survey. Springer, Berlin Heidelberg New York
17. Bander H, Nather W (1992) Fuzzy data analysis. Kluwer, Dordrecht