Effectiveness of Domain-Specific Cluster Representatives for Graphical Plots

Aparna S. Varde[1,2], Elke A. Rundensteiner[1], Carolina Ruiz[1], David C. Brown[1,3], Mohammmed Maniruzzaman[2,3] and Richard D. Sisson Jr.[2,3]
1. Department of Computer Science
2. Center for Heat Treating Excellence
3. Department of Mechanical Engineering
Worcester Polytechnic Institute, Worcester, MA, USA
(aparna|rundenst|ruiz|dcb|maniruzz|sisson)@wpi.edu *

ABSTRACT

Experimental results in scientific domains are often plotted as graphs of process variables. Clustering such graphs is useful for applications such as process comparison in which cluster representatives form at-a-glance depictions of each cluster. Randomly selected representatives are seldom effective in incorporating semantics, user interests and ease of interpretation. Hence there is a need to design domain-specific representatives. Users cannot always evaluate the design due to lack of time, huge volumes of experimental data and vague notions of effectiveness. Hence there is a need to automate evaluations using objective effectiveness measures. In this paper, we propose an approach called DesGraph that designs and evaluates domain-specific cluster representatives of graphs. Design is based on guided selection and construction giving candidates such as medoid and summarized graphs. An encoding analogous to the Minimum Description Length principle is proposed to assess the quality of the candidates taking into account complexity, information loss and user interests. The winning candidate is the one with the lowest encoding. DesGraph is experimentally evaluated in Materials Science. Designed representatives are found to consistently outperform random ones. Various representatives (e.g., medoid, summarized) are winners in different targeted applications.

1. INTRODUCTION

This paper addresses a sub-problem of clustering [7], namely, designing and evaluating domain-specific cluster representatives. We cluster *graphical plots of scientific functions* referred to in our work as *graphs* [13]. The graphs, resulting from experimental processes, plot a dependent versus an independent variable showing behavior of process parameters in a given domain. The domain of focus in this paper is the Heat Treating of Materials [2] that motivated this research. In applications such as process comparison, computational estimation and simulation [10, 13, 8] it is found useful to analyze the behavior of processes based on the similarity of their results [2]. This is facilitated by clustering graphs obtained from the corresponding processes. Cluster representatives serve as at-a-glance visual depictions of each cluster in such applications. However, a randomly selected representative may not incorporate all the necessary information about the cluster. This problem has been observed in our earlier work, AutoDomainMine, that uses cluster representatives for computational estimation [13, 14]. While displaying all the relevant information it is also important to avoid visual clutter. Moreover, different categories of users may be interested in visualizing different levels of detail. Given such considerations it is important to *design* domain-specific cluster representatives for targeted applications.

Having designed the representatives it is essential to assess their quality based on how well they capture information, avoid clutter and cater to user interests. Users are not always available to perform evaluations, especially at intermediate stages of system development. Also, huge volumes of data make it infeasible for users to conduct manual evaluations. Moreover, criteria for quality are seldom precisely defined. Users may have vague subjective notions of their expectations from a system. Hence it is desirable to automate the evaluations using objective measures of effectiveness.

In this paper we propose an approach called DesGraph that designs and evaluates domain-specific cluster representatives of graphical plots. In DesGraph we utilize two design methods, namely, guided selection and construction. In guided selection, the representative is chosen to be one object of the cluster, e.g., the graph that forms the cluster medoid [7]. In construction, the representative is a new object developed using cluster information, e.g., by superimposing all graphs in the cluster. These selected and constructed objects form candidate representatives in DesGraph. An effectiveness measure for evaluating these representatives is proposed in this paper. The proposed measure called the DesGraph Encoding is analogous to the Minimum Description Length [11] principle. The DesGraph Encoding incorporates the complexity of the cluster representative, information loss due to the representative and interests of

*This work is supported by the Center for Heat Treating Excellence (CHTE) and its member companies, and by the Department of Energy - Industrial Technology Program (DOE-ITP) Award Number DE-FC-07-01ID14197.

targeted users. Candidate representatives are compared using this encoding. The candidate with the lowest encoding is the winner.

DesGraph is experimentally evaluated using graphs in the Heat Treating domain [2]. On comparing designed representatives with random ones using the DesGraph Encoding, designed representatives are shown to be superior in quality. Different candidates such as medoid and summarized graphs are winners in different situations depending on factors such as user interests. The winners are useful in designing applications. For example in a given system, if ease of interpretation is more important than amount of information conveyed, this fact can be coded in the DesGraph Encoding. If the winner for the given data set then happens to be the medoid graph, then this is used in the system. DesGraph is also evaluated by integrating it with the AutoDomainMine system [12] for computational estimation in Heat Treating. Estimation using designed representatives is compared with the estimation in an earlier version of the system using random representatives. It is observed that designed representatives enhance estimation accuracy.

The following contributions are made in this paper.

- Designing domain-specific cluster representatives by guided selection and construction.
- Proposing a domain-specific effectiveness measure to assess the quality of the representatives.
- Evaluating the usefulness of the designed representatives for computational estimation.
- Gauging the use of the designed representatives in other applications based on the effectiveness measure.

The rest of the paper is organized as follows. Section 2 presents an overview of Heat Treating as needed for the given problem. Section 3 explains the DesGraph approach. Section 4 describes its experimental evaluation. Section 5 summarizes related work. Section 6 gives the conclusions.

2. DOMAIN DESCRIPTION

Graphs in Heat Treating. Figure 1 shows a graph called a heat transfer curve. It depicts the heat transfer coefficient h versus temperature T of a material where the heat transfer coefficient measures the heat extraction capacity in a rapid cooling process called quenching [2]. Some regions on the graph are more significant than others because they correspond to physical phenomena in the domain. Boiling Point region BP shows the temperature of the part being reduced to the boiling point of the cooling medium. Leidenfrost Point LF denotes the breaking of the vapor blanket resulting in rapid cooling. Slow Cooling region SC is where the quenching process ends [2]. Maximum and minimum heat transfer regions MAX and MIN respectively are statistical distinguishing factors.

Distance Metrics for Comparison. Different metrics from the literature [6, 7] can be used to compare these graphs, e.g., Euclidean distance based on absolute position of points and statistical distances based on statistical distinguishing factors. In addition, we define *Critical Distances* [14] such as Leidenfrost distance and Boiling Point distance based on the respective critical regions. Our earlier work LearnMet [14] learns semantics-preserving distance metrics for graphs, where a LearnMet metric is of the form

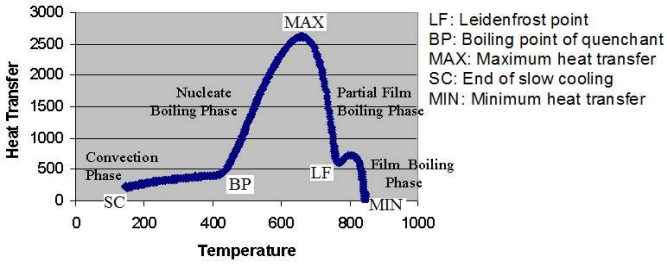

Figure 1: Heat Transfer Curve

$D = \Sigma_{i=1}^{m} w_i Dc_i$, each Dc_i being a component such as Euclidean, Statistical or Critical distance, w_i, its weight indicating relative importance, and m, the number of components. Metrics learned from LearnMet when used for clustering give higher clustering accuracy than individual metrics (e.g., Euclidean distance) when evaluated using true clusters over distinct test sets [14]. The output of LearnMet therefore is used as the notion of distance for the graphs.

3. THE DESGRAPH APPROACH

Given a set of clusters, DesGraph works as follows. It designs candidate representatives for each cluster by guided selection and construction, measures their effectiveness using the DesGraph Encoding based on MDL [11], and returns the designed representative as the winning candidate with the least encoding. If multiple candidates get equal values in the encoding then all are considered winners.

STEPS OF DESGRAPH

Input: Clusters of graphs, Notion of distance

1. Design candidate cluster representatives by
   (a) Guided selection as
      i. Nearest Representative
      ii. Medoid Representative
   (b) Construction as
      i. Summarized Representative
      ii. Combined Representative
2. Use *DesGraph Encoding* to measure effectiveness of candidates
3. Return designed representative (candidate with least encoding)

3.1 Candidates for Cluster Representatives

Consider the example of Cluster A in Figure 2. We explain design of candidate representatives based on this example.

3.1.1 Design by Guided Selection

In guided selection the representative is chosen as one of the objects of the cluster. Two candidate representatives, *nearest* and *medoid* are selected as shown in Figure 3.

Nearest Representative. The nearest representative is based on the concept of nearest neighbors using pairwise distances, as defined below.

$FOR\ f = 1\ to\ g$
$\quad SUM(f) = \Sigma_{i=1}^{g} D(G_f, G_i)$
$ENDFOR$
$RETURN\ R = G_f\ with\ lowest\ SUM(f)$

where G_f,G_i refer to individual graphs in the cluster, g is the

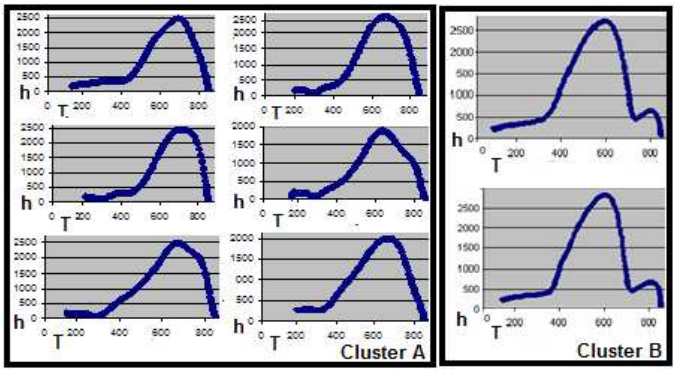

Figure 2: Clusters of Graphs

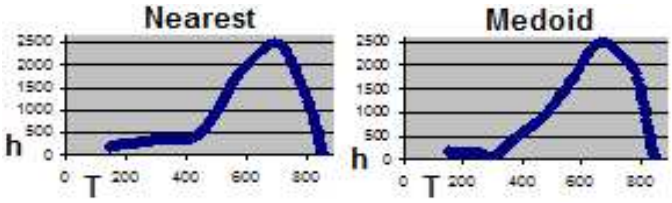

Figure 3: Selected Representatives

total number of graphs in the cluster, R is the representative graph and D is the distance between graphs using the given metric. We use sum and not sum of squares because the assumption is that squared distances are already incorporated in the metric. This representative, the nearest graph, shows users the member of the cluster that is nearest to the others using the given distance metric. Since the metric incorporates domain semantics this representative conveys nearness with respect to relative importance of regions on graphs.

Medoid Representative. A medoid representative, the graph in the cluster closest to its centroid, is defined below.

$FOR\ j = 1\ to\ n$
$\quad Cen(j) = \frac{1}{g}\Sigma_{i=1}^{g} G_i(j)$
$ENDFOR$
$FOR\ i = 1\ to\ g$
$\quad DIST(i) = \Sigma_{i=1}^{g} D(Cen, G_i)$
$ENDFOR$
$RETURN\ R = G_i\ with\ lowest\ DIST(i)$

where G_i refers to each graph, $G_i(j)$ is the value of the dependent variable (y-coordinate) at the j^{th} value of the independent variable (x-coordinate), n is the number of x-coordinates on the graphs, g is the number of graphs in the cluster, Cen is the cluster centroid and D is the distance using the given metric. The assumption is that the x-coordinates for all graphs are the same. Hence in computing the centroid, we take a mean of the y-coordinates only. This representative, the medoid graph, helps users visualize the object in the cluster closest to the average behavior of the dependent variable on the graphs.

3.1.2 Design by Construction

In construction the representative is an object developed using data in the cluster. We describe two such representatives, *summarized* and *combined* as shown in Figure 4.

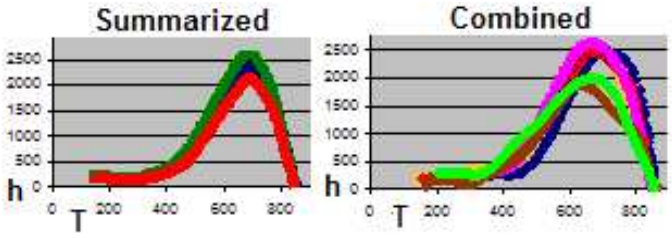

Figure 4: Constructed Representatives

Summarized Representative. The summarized representative presents a summary of information in the cluster. It is an average of graphs in the cluster with domain-specific upper and lower prediction limits. Average is computed as the cluster centroid while prediction limits are percentage upper and lower domain-specific thresholds added and subtracted from the average respectively, as follows.

$FOR\ j = 1\ to\ n$
$\quad R_{Av}(j) = \frac{1}{g}\Sigma_{i=1}^{g} G_i(j)$
$\quad R_{Up}(j) = R_{Av}(j) + \frac{U}{100} * R_{Av}(j)$
$\quad R_{Low}(j) = R_{Av}(X_j) - \frac{L}{100} * R_{Av}(j)$
$ENDFOR$
$RETURN\ R = R_{Up}, R_{Av}, R_{Low}$

where G_i refers to each graph, $G_i(j)$ is its y-coordinate at the j^{th} x-coordinate, n is the number of x-coordinates, g is the number of graphs in the cluster, R_{Av}, R_{Up} and R_{Low} are the average graph, upper limit and lower limit respectively, $R_{Av}(j)$, $R_{Up}(j)$ and $R_{Low}(j)$ being their respective y-coordinates at the j^{th} x-coordinate, U and L are percentage thresholds for upper and lower limits respectively, and R denotes the representative. Thresholds are obtained from a study of the data and discussions with experts. For example, in Heat Treating both thresholds are 10%. This representative, namely, the average graph with prediction limits, is a complex object consisting of 3 curves. It gives users a depiction of ranges of information in the cluster.

Combined Representative. The combined representative is constructed by superimposing all the graphs in a given cluster on each other as follows.

$FOR\ j = 1\ to\ n$
$\quad FOR\ i = 1\ to\ g$
$\quad\quad R_i = (G_i(j))$
$\quad ENDFOR$
$ENDFOR$
$RETURN\ R = R_i : i = 1\ to\ g$

where G_i is each graph, $G_i(j)$ is its y-coordinate at the j^{th} x-coordinate, n is the number of x-coordinates, g is the number of graphs in the cluster, and R is the representative. This representative, called the superimposed graph, is a complex object composed of g curves. It shows users the whole cluster with no information loss and depicts possible subtleties in the cluster. For example, the combined representative in Figure 4 shows that maximum heat transfer occurs at around the same temperature for all graphs in the cluster.

3.2 Effectiveness Measure for Representatives

We propose an effectiveness measure called the DesGraph Encoding for evaluating representative graphs. This encoding is analogous to the Minimum Description Length (MDL) principle [11]. MDL aims to minimize the sum of encoding

a theory and the examples using a theory. In DesGraph, the theory is the representative itself and the examples are all the objects in the cluster. However, the difference is that in DesGraph, we do not need to retrieve the original cluster from the encoding. Rather, we aim to compare the quality of the representatives in terms of how well they capture cluster information and how complex they are taking into account user interests. Hence the complexity of storing the representative graph and its distance from all graphs in the cluster are incorporated in the DesGraph Encoding. This encoding aims to minimize the sum of the number of bits to store the representative and the distance of all graphs from the representative. The user bias for complexity and distance is considered as percentage weights for each term. The encoding is given below.

The DesGraph Encoding:
$En_g = UBC * log_2(N_r) + UBD * log_2(\frac{1}{g}\Sigma_{i=1}^{g} D(R, G_i))$
where En_g = encoding for graphs
N_r = number of data points to store representative graph
R = the representative graph
G_i = each individual graph in the cluster
D = distance between graphs using the given metric
g = total number of graphs in the cluster
UBC = percentage weight giving user bias for complexity
UBD = percentage weight giving user bias for distance

The first term in the encoding, $log_2(N_r)$, is the complexity of storing the representative. Given that N is the number of x-coordinates, $N_r = N$ if R is *nearest* or *medoid*, $N_r = 3*N$ if R is summarized, and $N_r = g * N$ if R is combined.

The second term in the encoding, $log_2(\frac{1}{g}\Sigma_{I=1}^{g} D(R, G_i))$ is the average distance of each graph in the cluster from the representative. This distance gives the information loss with respect to domain semantics because it is computed using the given distance metric. Distance is calculated as $D(R, G_i)$ if R is *nearest* or *medoid*, as the minimum of $(D(R_{Av}, G_i)$,$D(R_{Up}, G_i)$ and $D(R_{Low}, G_i)$ if R is *summarized* and as the minimum of all values $D(R_i, G_i) : i = 1$ *to* g for the given G_i if R is *combined*.

Percentage weights UBC and UBD give user bias for complexity and distance terms in the encoding respectively. Default weights are 50% each, indicating equal importance of both terms. In some situations users are interested in capturing more information in the cluster and do not care about how complex the representative is. Thus complexity gets a lower weight. Some categories of users give high importance to complexity for reasons such as storage and ease of display. Hence complexity gets a higher weight.

Figure 5 shows calculations for measuring the effectiveness of representatives for Cluster A. Designed candidates are compared with each other and with a random representative. Complexity and Distance columns in the figure show values of the respective terms in the encoding without user bias. Columns (10/90), (50/50) and (90/10) give user bias for complexity and distance respectively. Winners for each column are shown in italics.

4. EXPERIMENTAL EVALUATION

DesGraph is implemented in Java and is experimentally evaluated using real data from Heat Treating. Standalone evaluation of DesGraph is performed using the proposed effectiveness measure. Additional evaluation is conducted by integrating DesGraph with the AutoDomainMine system for computational estimation [12]. The platform used for all the evaluations is a Mobile Intel Celeron (R) PC with a CPU Speed of 2 GHz, 192 MB of RAM and the Microsoft Windows XP Professional Version 2002 operating system.

Representative	Complexity	Distance	(10/90)	(50/50)	(90/10)
Nearest	11.96578428	5.623516	625.77	879.47	*1133.2*
Medoid	11.96578428	5.693533	632.08	882.97	1133.9
Summarized	13.55074679	5.786204	656.27	966.85	1277.4
Combined	16.13570929	0	*161.36*	*806.79*	1452.2
Random	11.96578428	6.376806	693.57	917.13	1140.7

Figure 5: Effectiveness of Representatives

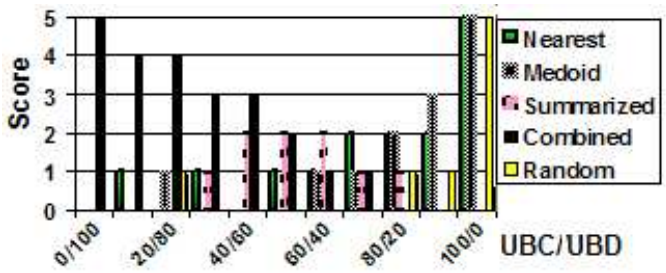

Figure 6: Statistics for Small Data Set

4.1 Standalone Evaluation of DesGraph

4.1.1 Evaluation Process

In the standalone evaluation clusters of graphs over different data sets are sent as input to DesGraph. Input parameters altered are weights of complexity and distance, data set size, number of clusters, and clustering seeds. The clustering algorithm used is k-means [7]. Output of DesGraph is the winning candidate for each cluster.

For comparison, a random representative is considered per cluster in the evaluation process. Scores are then assigned to each representative as the number of clusters in the data set in which it is the winner. For example, in a data set of 25 graphs in 5 clusters with (50/50) weights, if the winner is medoid for two clusters and combined for three, then the scores are, Nearest:0, Medoid:2, Summarized:0, Combined:3 and Random:0. The statistics is reported accordingly.

4.1.2 Evaluation Results

A summary of the evaluation of DesGraph in Heat Treating is presented here. We show the results of 330 experiments run with a small data set of 25 graphs in 5 clusters, a medium data set of 150 graphs in 10 clusters and a large data set of 400 graphs in 20 clusters. For each data set, user bias for complexity and distance is altered from (0/100) to (100/0) respectively in steps of 10. Each experiment is run 10 times, altering clustering seeds to build the clusters input to DesGraph. The average of 10 experiments is shown here. Statistics is reported as scores for representatives in Figures 6, 7 and 8 respectively. The observations made from the evaluation results are given below, followed by a discussion on their usefulness with respect to targeted applications.

Observations from Evaluation Results:

- For the small data set, combined representatives are

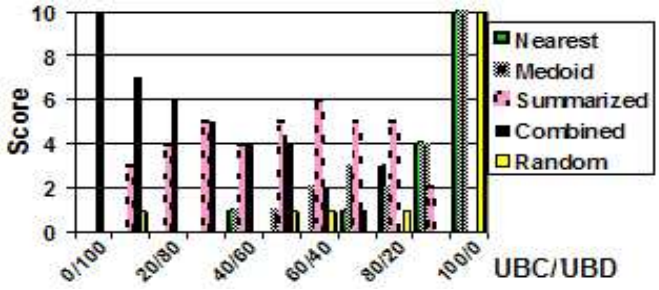

Figure 7: Statistics for Medium Data Set

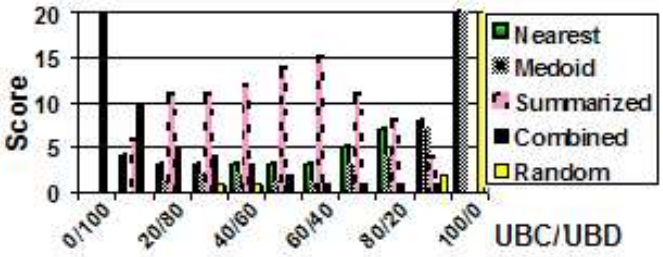

Figure 8: Statistics for Large Data Set

often winners followed by nearest and medoid.

- For the medium data set, the winners are usually summarized and combined representatives.
- For the large data set, summarized representatives are winners in most cases.
- For (10/90) weights, combined representatives win regardless of data set size.
- For (50/50) weights, summarized representatives win (with or without a tie) for all data sets.
- For (90/10) weights, all data sets have nearest/medoid representatives as winners.
- Random representatives lose almost always, except when users give zero weight to the distance term.

4.1.3 Discussion on Evaluation

These observations help design representatives in domain-specific applications as follows.

- The (90/10) weights are likely to arise in applications such as parameter selection [10]. Here a representative is used to study the behavior of a cluster to compare processes for selecting process parameters in industry. Hence a simple representative is desirable. Thus nearest/medoid representatives are useful, especially for large data sets.
- The (50/50) weights are typically found in simulation applications [8]. Users run simulations with representatives depicting ranges of information in the cluster. Hence the distance term matters because it denotes information loss. Complexity matters because simulations are time-consuming. Hence summarized representatives are useful for most data set sizes.

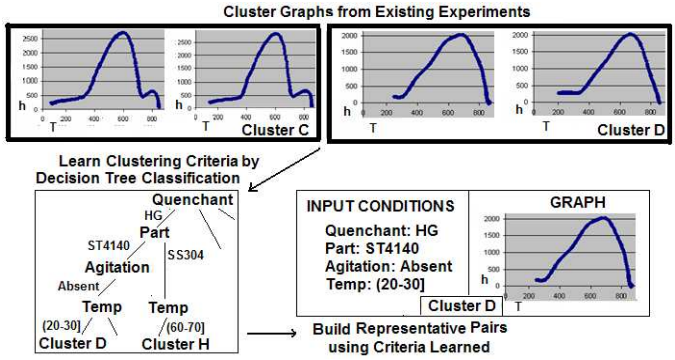

Figure 9: AutoDomainMine Estimation Approach

- The (10/90) weights would probably occur in applications such as decision support systems [15] for experts. In such systems it is important to study all information in the cluster to analyze process behavior in detail. Complexity of the representative does not matter as much. Thus combined representatives are useful in such applications.

4.2 Evaluation with System Integration

4.2.1 The AutoDomainMine System

DesGraph is developed for our larger project, a computational estimation system called AutoDomainMine [13, 14, 12]. AutoDomainMine estimates the graph obtained in a scientific experiment given its input conditions. The estimation approach first clusters graphs from existing experiments, then uses decision tree classifiers to learn relative importance of clustering criteria, and builds a representative pair of input conditions and graph per cluster. Representative pairs are the basis for estimating graphs of new experiments given their input conditions. The AutoDomainMine estimation approach is depicted in Figure 9.

4.2.2 Evaluation of DesGraph with AutoDomainMine

DesGraph is evaluated with AutoDomainMine to assess the effectiveness of the designed representatives. The intermediate stage of AutoDomainMine uses semantics-preserving distance metrics for clustering and randomly selected cluster representatives for estimation [14]. This stage is compared with the final stage that incorporates DesGraph using designed representatives [12]. The estimation in both stages is compared with the results of laboratory experiments from a distinct test set not used for training the technique. The difference in estimation accuracy with random and designed representatives shows the effectiveness of DesGraph. Estimation accuracy is computed as follows. Input conditions of each laboratory experiment in the test set are submitted to AutoDomainMine. The estimated graph is compared with the real graph from the laboratory experiment. If the two match within a domain-specific threshold then the estimation is considered accurate. Accuracy is then reported as the percentage of accurate estimations over the test set.

Observations from AutoDomainMine evaluation with and without DesGraph are shown in Figure 10. The designed representatives N, M, S and C correspond to the final stage of AutoDomainMine [12] while the random representative R refers to the intermediate stage [14]. It is observed

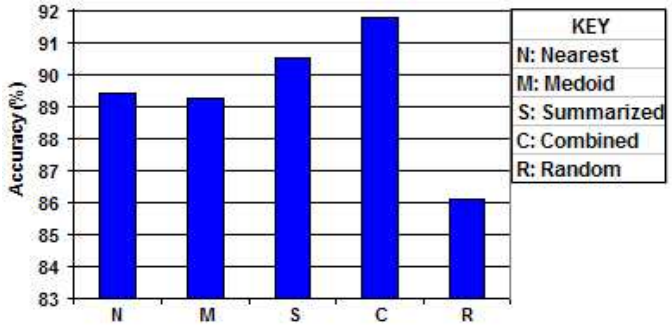

Figure 10: AutoDomainMine Evaluation Results

that designed representatives give higher estimation accuracy. Since the estimation provided by AutoDomainMine has various applications such as parameter selection, simulation tools and decision support [10, 8, 15], all the designed representatives found to be are useful.

5. RELATED WORK

Brecheisen et al [3] use reachability plots to extract significant clusters in a hierarchical cluster representation with medoids. Their restriction is that the representative must be an object of the cluster. We do not have such a restriction and can thus consider multiple design strategies.

Representatives for web information are built by manual intervention in [4] using image rating criteria of color, texture etc. No evaluation metrics are proposed. In our problem manual intervention for design is not feasible. Also, our evaluation criteria are different such as information loss (with respect to domain semantics), simplicity of interpretation and user interests in specific applications.

Janecek et al [5] address searches based on Semantic Fish Eye Views to display many large images in a small space. This is worthwhile when display space is critical which is not an issue in our targeted applications.

MDL encodings for cluster evaluation are proposed in the literature, e.g., [1, 9]. However, these encodings are not used to evaluate different types of representatives. Moreover, they need to retrieve the original cluster from the encoding which is not a requirement in our problem.

6. CONCLUSIONS

This paper proposes an approach called DesGraph for designing and evaluating domain-specific cluster representatives of graphical plots. Design involves guided selection and construction giving candidate representatives such as nearest and superimposed graphs. An effectiveness measure called DesGraph Encoding, analogous to MDL, is proposed for assessing the quality of the representatives. Experimental evaluation in Heat Treating shows that designed cluster representatives have distinctly better quality than random ones. Different types of designed representatives are found to be suitable in different targeted applications.

Ongoing work in this area involves designing and evaluating and domain-specific cluster representatives for sets of experimental input conditions leading to the graphs. This is another sub-problem being addressed as a part of the AutoDomainMine project. The designed representatives of conditions along with those of graphs will be used to build representative pairs for estimation in AutoDomainMine. The use of the designed representatives of conditions in other targeted applications will also be assessed.

7. REFERENCES

[1] Banerjee, A. and Langford J.: An Objective Evaluation Criterion for Clustering. KDD (Aug 2004) pp. 515-520.

[2] Boyer, H. and Cary, P.: Quenching and Control of Distortion. ASM International (1989).

[3] Brecheisen, S., Kriegel, H., Kroger, P., Pfeifle, M. and Viermetz, M.: Representatives for Visually Analyzing Cluster Hierarchies. KDD MDM (Aug 2003) pp. 64-71.

[4] Helfman, J. and Hollan, J.: Image Representations for Accessing and Organizing Web Information. SPIE II (2001) pp. 91-101.

[5] Janecek, P., and Pu, P.: Opportunistic Search with Semantic Fisheye Views. Swiss Federal Institute of Technology, Lausanne (2004) TR IC/2004/42.

[6] Keim, D. and Bustos, B.: Similarity Search in Multimedia Databases. ICDE (Mar 2004) pp. 873-874.

[7] Kaufman, L., and Rousseeuw, P.: Finding Groups in Data: Introduction to Cluster Analysis. John Wiley (1988).

[8] Lu, Q., Vader, R., Kang, J. and Rong, Y.: Development of a Computer-Aided Heat Treatment Planning System. Heat Treatment of Metals (2002) Vol. 3, pp. 65-70.

[9] Lent, B., Swami, A., and Widom, J.: Clustering Association Rules. ICDE (Apr 1997) pp 220-231.

[10] Maniruzzaman, M., Chaves, J., McGee, C., Ma, S. and Sisson, R.: The CHTE Quench Probe System. In ICFDM (Jul 2002) pp 13-17.

[11] Rissanen, J.: Stochastic Complexity and the MDL Principle. In Econometric Reviews Vol. 6 (1987) pp. 85-102.

[12] Varde, A., Rundensteiner E., Ruiz, C., Brown, D., Maniruzzaman, M. and Sisson, R.: Integrating Clustering and Classification for Estimating Process Variables in Materials Science. AAAI Symposium (Jul 2006) submitted Apr 2006.

[13] Varde, A., Rundensteiner, E., Ruiz, C., Maniruzzaman, M. and Sisson, R.: Data Mining over Graphical Results of Experiments with Domain Semantics. ACM ICICIS (Mar 2005) pp. 603-611.

[14] Varde, A., Rundensteiner E., Ruiz, C., Maniruzzaman, M. and Sisson, R.: Learning Semantics-Preserving Distance Metrics for Clustering Graphical Data. KDD MDM (Aug 2005) pp. 107-112.

[15] Varde, A., Takahashi, M., Rundensteiner E., Ward, M., Maniruzzaman, M. and Sisson, R.: QuenchMiner^{TM}: Decision Support for Optimization of Heat Treating Processes. IEEE's IICAI (Dec 2003) pp. 993 - 1003.