Comparing Mathematical and Heuristic Approaches for Scientific Data Analysis

Aparna S. Varde [1,2], Shuhui Ma [2,3], Mohammed Maniruzzaman [2,3],
Elke A. Rundensteiner [1], David C. Brown [1,3] 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 01609, USA
(aparna | mashuhui | maniruzz | rundenst | dcb | sisson)@wpi.edu

Abstract

Scientific data is often analyzed in the context of domain-specific problems, e.g., failure diagnostics, predictive analysis and computational estimation. These problems can be solved using approaches such as mathematical models or heuristic methods. In this paper we compare a heuristic approach based on mining stored data with a mathematical approach based on applying state-of-the-art formulae to solve an estimation problem. The goal is to estimate results of scientific experiments given their input conditions. We present a comparative study based on sample space, time complexity and data storage with respect to a real application in Materials Science. Performance evaluation with real Materials Science data is also presented, taking into account efficiency and accuracy. We find that both approaches have their pros and cons in computational estimation. Similar arguments can be applied to other scientific problems such as failure diagnostics and predictive analysis. In the estimation problem in this paper, heuristic methods outperform mathematical models.

1. Introduction

Scientific data in domains such as Materials Science is often analyzed in the context of domain-specific applications. An example is computational estimation [20] where the results of experiments are estimated without conducting real experiments in a laboratory. Another application is failure diagnostics [17] where existing cases are used to diagnose causes of failures such as distortion in materials. A related application is predictive analysis [21] where process variables are predicted apriori to assist parameter selection so as to optimize the real processes.

This paper describes the use of mathematical and heuristic approaches in such scientific data analysis. The goal is to perform a comparative study between these two approaches. We focus on a domain-specific computational estimation [20] problem and present a detailed study of mathematical and heuristic solution approaches. The domain of focus is Heat Treating of Materials [16]. The result of a heat treating experiment is plotted as a heat transfer curve [16]. Scientists are interested in estimating this curve given experimental input conditions.

Mathematical models for estimation are based on formulae derived from theoretical calculations [2, 16]. They provide definite solutions under certain situations. However, existing mathematical models are often inapplicable under certain circumstances [9, 10]. For example, in Heat Treating there is a direct-inverse heat conduction model for estimating heat transfer curves [2]. However, if the real experiment is not conducted, this model requires initial time-temperature inputs to be given by domain experts each time the estimation is performed. This is not always possible [10].

Heuristic methods are often based on approximation. A heuristic by definition is a rule of thumb likely to lead to the right answer but not guaranteed to succeed [15]. However heuristic methods are applicable in some situations where mathematical models cannot be used or do not provide adequate solutions. In our earlier work [20], we have proposed a heuristic approach based on integrating the data mining techniques of clustering and classification as a solution to a computational estimation problem. When applied to estimating heat transfer curves, this approach works well in many situations where mathematical models in heat treatment are not feasible.

In this paper, we present a comparative study between mathematical and heuristic approaches in estimation taking into account sample space, time complexity and data storage. Sample space refers to the number of experiments that can be estimated under various conditions. Time complexity refers to the computation of the mathematical models or heuristic methods are in terms of execution time. Data storage refers to the amount of data stored in the database in each approach.

We also provide performance evaluation with real data from the Heat Treating domain considering efficiency and accuracy. The efficiency of the approach relates to how fast it can perform the estimation. The accuracy of the estimated results refers to how close the estimation is to the result of a real laboratory experiment.

It is found that both mathematical and heuristic approaches have their advantages and disadvantages. For the given estimation problem in this paper, we find that heuristic methods are generally better than existing mathematical models.

The arguments made for computational estimation can also be considered valid in the context of the other applications such as failure diagnostics [17] and predictive analysis [21]. Detailed discussion on each of these is beyond the scope of this paper.

The following contributions are made in this work:

- Description of mathematical and heuristic approaches in computational estimation.
- Comparative study on sample space, time complexity and data storage.
- Performance evaluation with real data from Materials Science.

The rest of this paper is organized as follows. Section 2 gives a background of the computational estimation. Sections 3 and 4 describe mathematical and heuristic solutions to this problem respectively. Sections 5 and 6 give the comparative study and performance evaluation respectively. Section 7 outlines related work. Section 8 gives the conclusions.

2. Computational Estimation Problem

In scientific domains such as Materials Science and Mechanical Engineering experiments are performed in the laboratory with specified input conditions and the results are often plotted as graphs. The term graph in this paper refers to a two-dimensional plot of a dependent versus an independent variable depicting the behavior of process parameters. These graphs serve as good visual tools for analysis and comparison of the processes. Performing real laboratory experiments and plotting such graphs consumes significant time and resources, motivating the need for computational estimation.

We explain this with an example from the domain of Heat Treating of Materials [16] that inspired this work. Heat treating is a field in Materials Science that involves the controlled heating and rapid cooling of a material in a liquid or gas medium to achieve desired mechanical and thermal properties [16].

Figure 1 shows an example of the input conditions and graph in a laboratory experiment in *quenching*, namely, the rapid cooling step in heat treatment. The quenchant name refers to the cooling medium used, e.g., T7A, HoughtoQuenchG. The part material incorporates the characteristics of the part such as its alloy content and composition, e.g., ST4140, Inconel600. The part may have a thick or thin oxide layer on its surface. A sample of the part called the probe is used for quenching and has certain shape and dimensions characterized by the probe type. During quenching, the quenchant is maintained at a given temperature and may be subjected to a certain level of agitation, i.e., high or low. All these parameters are recorded as input conditions of the quenching experiment.

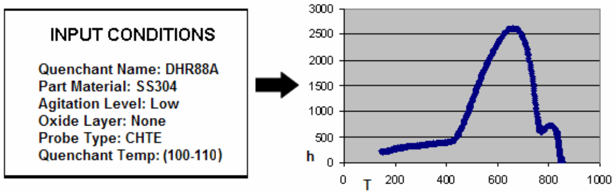

Figure 1: Example of Input Conditions and Graph

The result of the experiment is plotted as a graph called a heat transfer coefficient curve. This depicts the heat transfer coefficient *h* versus part temperature *T*. The heat transfer coefficient measures the heat extraction capacity of the process and depends on the cooling rate and other parameters such as part density, specific heat, area and volume [2, 9]. The heat transfer curve characterizes the experiment by representing how the material reacts to rapid cooling.

Materials scientists are interested in analyzing this graph to assist decision-making about corresponding processes. For instance, for the material *ST4140*, a kind of steel, heat transfer coefficient curves with steep slopes imply fast heat extraction capacity. The corresponding input conditions could be used to treat this steel in an application that requires such a capacity [22].

However, performing such an experiment in the laboratory takes 5 to 6 hours and the resources require a capital investment of thousands of dollars and recurring costs worth hundreds of dollars [10, 20].

It is thus desirable to computationally estimate in an experiment the resulting graph given the input conditions. The estimation problem is as follows [22]:

- Given: The input conditions of a scientific experiment
- Estimate: The resulting graph depicting the output of the experiment

We describe the solutions to this estimation problem with mathematical and heuristic approaches.

3. Mathematical Modeling Approach

Mathematical models are based on theoretical formulae that are often derived in a domain-specific manner. We explain mathematical modeling with reference to the problem of estimating heat transfer curves. This problem translates to estimating heat transfer coefficients as a function of temperature.

The estimation method presented here is based on the extension of the sequential function specification method of Beck et al [2]. It uses state-of-the-art formulae for heat transfer [16]. The mathematical model relates to direct and inverse heat conduction [9, 10].

3.1. Direct Heat Conduction

The mathematical formulation of the direct heat conduction problem when the surface heat flux is considered known is given by [10]:

$$\rho C_p(T)\frac{\partial T}{\partial t} = \frac{1}{r}\frac{\partial}{\partial r}\left(rk(T)\frac{\partial T}{\partial r}\right) \qquad (1a)$$

where,

k = thermal conductivity of the probe

C_p = specific heat of the probe

ρ = density of the probe

Boundary conditions are:

surface heat flux, $$q_b = h_b\left(T - T_f\right)\Big|_{r=R} = -k\frac{\partial T}{\partial r}\Big|_{r=R} \qquad (1b)$$

$$\frac{\partial T}{\partial r}\Big|_{r=0} = 0 \qquad (1c)$$

and the initial condition is:

$$T(r,0) = T_o \qquad (1d)$$

where,

q = surface heat flux

h_b = surface heat transfer coefficients

$Y(t)$ = measured temperature at center of probe

R = radius of the probe.

This direct problem can readily be solved by classical solutions or numerical solution techniques [16].

3.2. Inverse Heat Conduction

The mathematical formulation of the inverse heat conduction problem is given by [10]:

$$\rho C_p(T)\frac{\partial T}{\partial t} = \frac{1}{r}\frac{\partial}{\partial r}\left(rk(T)\frac{\partial T}{\partial r}\right) \qquad (2a)$$

where,

k = thermal conductivity of the probe

C_p = specific heat of the probe

ρ = density of the probe

Boundary conditions are:

surface heat flux, $$q_b = h_b\left(T - T_f\right)\Big|_{r=R} = -k\frac{\partial T}{\partial r}\Big|_{r=R} \qquad (2b)$$

$$\frac{\partial T}{\partial r}\Big|_{r=0} = 0 \qquad (2c)$$

and, initial condition is:

$$T(r,0) = T_o \qquad (2d)$$

where the surface heat flux q_b is unknown; temperature measurements are considered to be taken

$$T(0,t_j) = Y_j \quad j - 1,2,...,N \qquad (2e)$$

with a single sensor placed at $r = 0$ at time t_j are given over the whole time domain $0 < t \le t_f$, where t_f is the final measurement time.

Then the inverse problem can be stated as follows: By utilizing the N measured data $Y_j\left(j = 1,2,...,N\right)$, estimate the N heat flux components $q\left(t_j\right) \equiv q_j\left(j = 1,2,...,N\right)$

3.3. Steepest Descent Method for Estimation

Using the direct and inverse heat conduction equations, heat transfer coefficients are estimated using the Steepest Descent Method [9]. In this method, initial heat transfer coefficients values are given as inputs. Using these, the method works as follows.

(i) Accept the given heat transfer coefficients.

(ii) Use the heat transfer coefficients in the direct heat conduction equation to obtain heat flux values.

(iii) Substitute the heat flux values in the inverse heat conduction method.

(iv) Calculate the heat transfer coefficients using these heat flux values.

(v) If error between heat transfer coefficients in (iv) and (i) is minimal or maximum number of iterations is reached then stop. Output heat transfer coefficients in (iv) as the estimated heat transfer coefficients.

(vi) Else go to step (i) using heat transfer coefficients calculated in step (iv).

In order to use this model, the initial heat transfer coefficients need to be given. These are calculated based on time-temperature data using the input conditions of the experiment. From time-temperature data, initial heat transfer coefficients are obtained using state-of-the-art formulae [2, 9, 16]. However, since actual measurements are not taken as in equation (2e) by performing real experiments the time-temperature inputs must be supplied by experts each time the estimation is performed. The experts usually guess these inputs based on the experimental input conditions such as quenchant and part.

Thus, heat transfer coefficients can be estimated mathematically by direct and inverse heat conduction using the steepest gradient descent method.

4. Heuristic Approach based on Data Mining

The term *heuristic* originates from the Greek word “heureskein” meaning “to find” or “to discover” [15]. As stated by Newell et al “A process that may solve a given problem but offers no guarantees of doing so is called a heuristic for that problem” [11]. Nevertheless, heuristic methods in the literature often provide good solutions to many problems [15].

We have proposed a heuristic approach called AutoDomainMine [20] to solve the given computational estimation problem. The assumption in this approach is that data obtained from existing experiments is stored in a database and is available for analysis.

4.1. The AutoDomainMine Approach

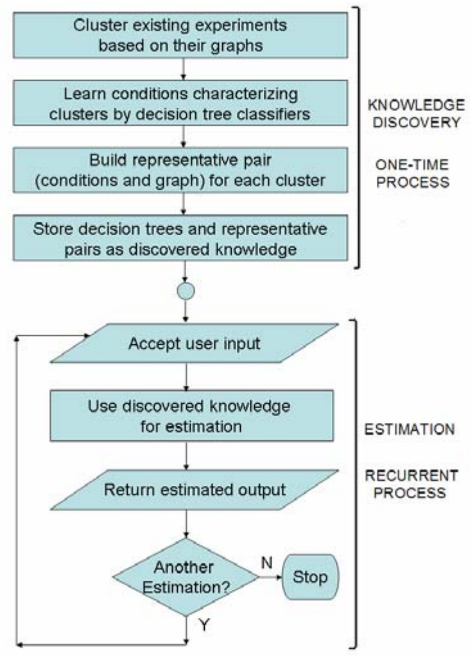
Cluster existing experiments based on their graphs
Learn conditions characterizing clusters by decision tree classifiers
Build representative pair (conditions and graph) for each cluster
Store decision trees and representative pairs as discovered knowledge
KNOWLEDGE DISCOVERY
ONE-TIME PROCESS
Accept user input
Use discovered knowledge for estimation
Return estimated output
Another Estimation?
N
Stop
Y
ESTIMATION
RECURRENT PROCESS

Figure 2: The AutoDomainMine Approach

AutoDomainMine [20] involves a one-time process of knowledge discovery from previously stored data and a recurrent process of using the discovered knowledge for estimation. This approach is illustrated in Figure 2.

AutoDomainMine discovers knowledge from existing experimental data by integrating the two data mining techniques of clustering and classification. Clustering is the process of placing a set of objects into groups of similar objects [4, 7]. Classification is a form of data analysis that can be used to extract models to predict categories [4, 8]. These two data mining techniques are integrated for knowledge discovery as follows.

4.2. Knowledge Discovery in AutoDomainMine

The knowledge discovery process is shown in Figure 3. Clustering is first done over the graphs obtained from existing experiments. We use a suitable algorithm such as k-means [7] with a domain-specific distance metric as the notion of distance [22]. Once the clusters of experiments are identified, the clustering criteria, namely, the input conditions that characterize each cluster are learned by decision tree classification [8]. This helps understand the relative importance of conditions in clustering. The decision tree paths and the clusters they lead to are used to design a representative pair of input conditions and graph per cluster so as to preserve domain semantics [22]. The decision trees and representative pairs form the discovered knowledge used for estimation as follows.

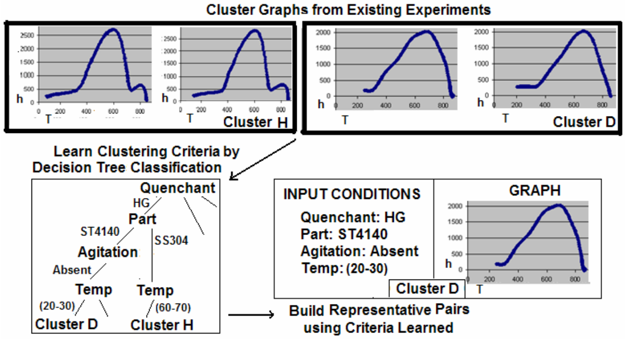

Figure 3: AutoDomainMine - Knowledge Discovery

4.3. Estimation in AutoDomainMine

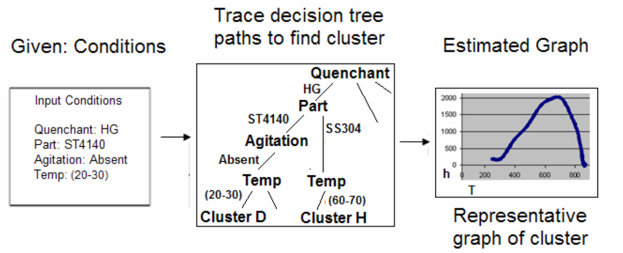

Figure 4: AutoDomainMine – Estimation

The process of estimation is shown in Figure 4. In order to estimate a graph, given a new set of input conditions, the decision tree is searched to find the closest matching cluster. The representative graph of that cluster is the estimated graph for the given set of conditions. If a complete match cannot be found then partial matching is done based on the higher levels of the tree using a domain-specific threshold [22]. Note that this estimation incorporates the relative importance of conditions identified by the decision tree.

5. Comparative Study

We compare the mathematical and heuristic approaches based on sample space, time complexity and data storage.

5.1. Sample Space

The sample space of any estimation problem is the number of cases it can estimate [15]. We explain the calculation of sample space with reference to the estimation problem in this paper.

Sample Space Calculation: The sample space is calculated as a product of the number of possible values of each experimental input condition. Each input condition is described by an attribute that gives its name and a value that gives its content [22].

Thus we have, sample space $S = \prod_{c=1}^{A} V_c$ (3) where,

A = total number of attributes (conditions)
V_c = number of possible values of the condition

Consider the example of estimating heat transfer curves. In this example, the input conditions are:

- Quenchant Name: T7A, DurixolV35 etc.
- Part Material: ST4140, SS304 etc.
- Agitation Level: Absent, High, Low
- Oxide Layer: None, Thin, Thick
- Probe Type: CHTE, IVF etc.
- Quenchant Temperature: 0 to 200 C

The number of possible values of each of these is:

- Quenchant Name: 9 values
- Part Material: 4 values
- Agitation Level: 3 values
- Oxide Layer: 3 values
- Probe Type: 2 values.
- Quenchant Temperature: 20 ranges

The sample space is given by a product of these values. Hence, in this example we have:

$$\textit{Sample Space} = 9 \times 4 \times 3 \times 3 \times 2 \times 20 = 12960$$

We now discuss this with reference to our mathematical and heuristic approaches.

5.1.1. Mathematical Approach. In this approach, the estimation of heat transfer coefficients is performed using the direct and inverse heat conduction equations [9, 10]. However, in order to apply these equations, data on time and temperature is needed. If the real laboratory experiment is not conducted then this data is typically supplied by domain experts.

Thus, in this process domain expert intervention is needed each time the estimation is performed. Thus, in order to cover a sample space of 12960 experiments, the domain experts would need to provide the time-temperature inputs 12960 times which seems rather infeasible. Besides the fact that supplying these inputs is time-consuming and cumbersome, it is not always possible for the experts to guess them based on experimental input conditions. This is a major drawback of the mathematical approach related to sample space.

However, an advantage of this approach is that no other data on previous experiments needs to be stored in advance to cover this sample space. The state-of-the-art formulae can be directly applied.

This advantage and disadvantage is further clarified as we discuss the heuristic solution.

5.1.2. Heuristic Approach. The heuristic solution approach to our estimation problem is AutoDomainMine [20]. In this approach, when the input conditions of a new experiment are submitted, the decision tree paths are traced to find the closest match. The representative graph of the corresponding cluster is conveyed as the estimated result. When an exact match is not found, a partial match is conveyed using higher levels of the tree. Thus, even if data on all the possible combinations of inputs is not available, an approximate answer can still be provided.

Hence, in order to cover the sample space of the estimation it is not necessary to supply time-temperature data for each new experiment whose results are to be estimated. The estimation can be performed simply by supplying the input conditions of the new experiment. Thus, the whole sample space of 12960 experiments can be covered without domain expert intervention each time the estimation is performed. This is an advantage of the heuristic approach with reference to sample space.

However, in order to perform the estimation in AutoDomainMine, data from existing laboratory experiments needs to be stored in the database. This forms the basis for knowledge discovery and estimation. This is seemingly a disadvantage of the heuristic approach. However, the amount of data from existing experiments can be much lower than the sample space.

For example, in Heat Treating the number of experiments stored is 500. With this, AutoDomainMine gives an accuracy of around 94% as elaborated later.

5.2. Time Complexity

The time complexity of any approach refers to the execution time of the technique used for computation [4].

5.2.1. Mathematical Approach. In the direct-inverse heat conduction mathematical model, the time complexity $t_M(E)$ of each estimation is given as [9]:

$$t_M(E) = O(n^2 \times i) \qquad (4)$$ where,

n = number of time-temperature data points supplied
i = number of iterations for convergence to minimal error

Each such data point corresponds to the measurement of heat transfer coefficient at one instance of time.

In the given problem the maximum number of data points supplied would be 1500 and the minimum number would be 25. On an average 100 data points are supplied. The number of iterations for convergence is typically of the order of 100 iterations [9].

Thus, we have the following time complexities.

Worst Case: $t_M(E) = O(1500^2 \times 100)$ (5a)

Average Case; $t_M(E) = O(100^2 \times 100)$ (5b)

Best Case: $t_M(E) = O(25^2 \times 100)$ (5c)

Since the data points need to be provided for each estimation, the time complexity t_M *(S*) over the whole sample space S is given by:

$$t_M(S) = S \times t_M(E) \qquad (6)$$ where,

$t_M(E)$ is the time complexity of each estimation.

Thus, we have the following time complexities over the whole sample space for the worst, average and best cases respectively.

Worst Case: $t_M(S) = S \times O(1500^2 \times 100)$ (7a)

Average Case: $t_M(S) = S \times O(100^2 \times 100)$ (7b)

Best Case: $t_M(S) = S \times O(25^2 \times 100)$ (7c)

Given a sample space of *S = 12960*, it is clear that these time complexities are huge.

5.2.2. Heuristic Approach. In the heuristic approach AutoDomainMine [20], the knowledge discovery process of clustering followed by classification is executed one-time, while the estimation process of searching the decision tree paths to find the closest match is recurrent. The complexities of each are calculated as follows.

Consider $t_H(D)$ to be the time complexity of the knowledge discovery process in the heuristic approach. This is calculated as the sun of the time complexities of the clustering and classification step respectively. We use k-means clustering [7] and decision tree classification with J4.8 [8]. The complexities of these respective algorithms [4] are used to compute the complexity of the knowledge discovery process in AutoDomainMine. Thus given that,

g = number of graphs (experiments) in database
k = number of clusters
i = number of iterations in the clustering algorithm
we have,

$$t_H(D) = t_H(Clustering) + t_H(Classification) \qquad (8a)$$

where, $t_H(Clustering) = O(gki)$ (8b)

and $t_H(Classification) = O(g\log(g))$ (8c)

Hence, $t_H(D) = O(gki) + O(g\log(g))$ (8d)

Now consider that the time complexity of each estimation in the heuristic approach is $t_H(E)$*.* The manner in which the estimation is performed in AutoDomainMine is by searching the decision tree paths to find the closest match with the given input conditions of a new experiment. From a study of the literature [4, 14, 15], we find that this search problem in general has a complexity of *O(log (N))* where *N* is the number of entries in the database from which the tree was generated. Thus, in our context this translates to *O(log (g))* since g = number of graphs in the database = number of experiments (i.e., database entries). Thus,

$$t_H(E) = O(\log(g)) \qquad (9)$$

Hence, given a sample space *S,* the time complexity $t_H(S)$ over the whole space is calculated as:

$$t_H(S) = t_H(D) + S \times t_H(E) \qquad (10)$$ where,

$t_H(D)$ = complexity of knowledge discovery (one-time)
$t_H(E)$ = complexity of each estimation (recurrent)
S = sample space

Thus, from the calculation of the time complexities $t_H(D)$ and $t_H(E)$ respectively, we get,

$$t_H(S) = O(gki) + O(g\log(g)) + S \times O(\log(g)) \qquad (11)$$ where,

g = number of graphs (experiments) in database
k = number of clusters
i = number of iterations in the clustering algorithm
S = sample space

Given this, we now consider the time complexities in the best, average and worst case in our problem.

Note that the maximum value of *g* is equal to all the experiments in the database, i.e. 500 in this context. The minimum value of *g* is empirically set to be at least 1/5 of the total number of experiments [22]. Thus, *g* is at least 100. The average value for *g* is considered to be half the total number of experiments, i.e., g is equal to 250 in the average case [22]. The number of clusters *k* is usually set close to the square root of the number of graphs *g* since this value is found to yield the highest classifier accuracy [22]. Thus, for *g = 500, k =22*; for *g = 250, k = 16;* and *for g = 100, k = 10.* The number of iterations in the clustering algorithm is typically of the order of 10. Given these values, we have the following time complexities in the worst, average and best cases respectively.

Worst: $t_H(S) = O(500 \times 22 \times 10) + O(500\log(500)) + S \times O(\log(500))$ (12a)

Avg: $t_H(S) = O(250 \times 16 \times 10) + O(250\log(250)) + S \times O(\log(250))$ (12b)

Best: $t_H(S) = O(100 \times 10 \times 10) + O(100\log(100)) + S \times O(\log(100))$ (12c)

These complexities in the heuristic approach are much lower than the worst, average and best case time complexities respectively in the mathematical modeling approach. This is an advantage of the heuristic method.

5.3. Data Storage

The data storage criterion refers to the quantity of data stored from existing experiments in each approach.

5.3.1. Mathematical Approach. This approach uses theoretical formulae and inputs supplied by domain experts each time the estimation is performed. No data from previously performed experiments is utilized in the computation. Hence, in theory the quantity of data stored for this approach is zero. Thus, given that *Q* refers to the quantity of data, we find that in the mathematical model, *Q= 0*. This is an advantage of the mathematical approach.

However, it is to be noted that the experts while providing initial time-temperature inputs to this model, may refer to existing experiments. Thus, in practice data stored from previously performed experiments could perhaps be useful in mathematical modeling. But this data storage is not a requirement of the model per se.

5.3.1. Heuristic Approach. This approach uses the existing experiments in the database for knowledge discovery and estimation. Given that g is the number of graphs (experiments) in the database, n is the number of data points stored per graph and A is the number of attributes stored for each experiment, the quantity Q of data stored in the heuristic approach is given as

$$Q = g \times n \times A$$

The heuristic approach cannot work without data from previous experiments. This is one of the situations where the mathematical model wins over the heuristic method.

Theoretically, there is no bound on the minimum quantity of data that needs to be stored in order to perform the estimation heuristically. However, the more the data from existing experiments, the more accurate is the estimation. This is because a greater number of experiments are available for knowledge discovery by clustering and classification and a greater number of decision tree paths can be searched for estimation. Also, the more distinct the input conditions are, the better it is for the heuristic approach. This is because a greater number of distinct paths can be identified in the decision tree to more classify new experiments.

Note that in scientific domains experiments are often designed using the Taguchi metrics [13]. The input conditions are selected such that 100 experiments can effectively represent approximately 300 experiments. This in turn enhances the sample space and accuracy of the estimation. It is therefore desirable that Taguchi metrics [13] be used for the experimental setup to provide effectively more data for the heuristic approach.

6. Performance Evaluation

6.1. Accuracy

Accuracy is a quality measure that refers to how close the estimated result is to the output of a real experiment. The evaluation of accuracy is explained with reference to the mathematical and heuristic approaches individually.

6.1.1 Mathematical Approach. The accuracy of mathematical models in Heat Treating is evaluated as follows [9]. The heat transfer curve estimated by a given mathematical model is superimposed over the real heat transfer curve obtained from a laboratory experiment conducted with the same input conditions. If the two match each other as per the satisfaction of the domain experts, then the estimation is considered to be accurate. We present a summary of the evaluation.

Test 1: This test is conducted with the inputs below.

- Quenchant Name: HoughtoQuenchG
- Part Material: ST4140
- Agitation Level: Low
- Oxide Layer: None
- Probe Type: CHTE
- Quenchant Temperature: 60 - 70 C

Figure 5 shows the heat transfer curves plotted as heat transfer coefficient h versus temperature T from both the real laboratory experiment and the mathematical model. According to the experts, the results show much difference in the magnitude of heat transfer coefficient as well as the temperature at which the maximum heat transfer coefficient occurs [9]. Moreover, the heat transfer curve from mathematical models shows the occurrence of a Leidenfrost point[1] (LF) [16] while the curve from the real experiment does not. Thus this estimation is considered inaccurate by the experts.

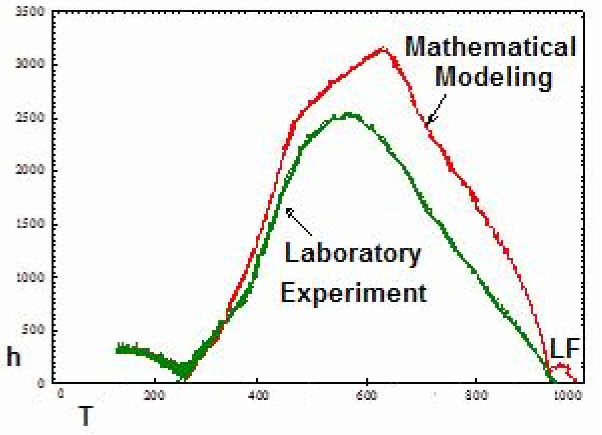

Figure 5: Output of Test 1 in Mathematical Approach

Test 2: The input conditions in this test are as follows.

- Quenchant Name: T7A
- Part Material: ST4140
- Agitation Level: High
- Oxide Layer: Thin
- Probe Type: CHTE
- Quenchant Temperature: 20 - 30 C

Figure 6 shows the output of this test in terms if the heat transfer curves obtained from the laboratory experiment and the mathematical model. Both the curves show the occurrence of the Leidenfrost point [16], which is one of the important parameters that characterize the quenching process. Moreover both curves have the Leidenfrost points occurring at approximately the same values of temperature and heat transfer. The difference between the maximum heat transfer of the two curves is also within acceptable limits with respect to temperature and heat transfer coefficient. Positions of most other

[1] The Leidenfrost point marks the breaking of a vapor blanket around a part. Heat transfer curves with and without a Leidenfrost point depict distinctly different cooling tendencies [16].

points on the two curves are also similar. Thus the experts conclude that this estimation is accurate

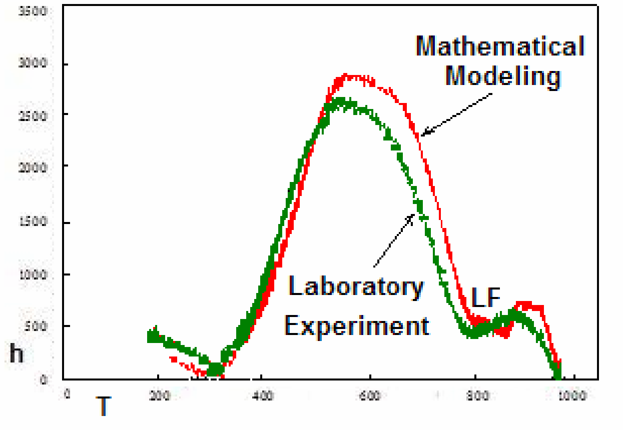

Figure 6: Output of Test 2 in Mathematical Approach

Likewise on conducting several tests with different input conditions, the estimation accuracy of the mathematical models is found to be in the range of approximately 85 to 90%. However, this is subject to the availability of good time-temperature inputs from experts.

6.1.1 Heuristic Approach. The accuracy of the heuristic model is evaluated with formal surveys conducted by the targeted users of the system [22]. The users run tests with the tool developed using the AutoDomainMine technique (a CHTE trademark). The N-holdout strategy [15] is used for evaluation. Among the 500 experiments in the Heat Treating database, 400 are used for training the technique and the remaining 100 are kept aside as the test set.

Tests are conducted as follows. In each test, the users enter the input conditions of a real experiment from the test set. They observe the estimated output of AutoDomainMine and compare it with the output of the corresponding real experiment. If the real and estimated results are close enough as per user satisfaction then the estimated is considered to be accurate. Accuracy is then reported as the percentage of accurate estimations over all the tests conducted [22].

We target the users of various applications of AutoDomainMine such as parameter selection [18], simulation tools [6], decision support [20] and intelligent tutors [3]. Accuracy is reported in the context of each application. A summary of our evaluation is presented.

Test 1: For the sake of comparison, this test is conducted with the same input conditions as Test 1 of the mathematical approach (see Section 6.1.1).

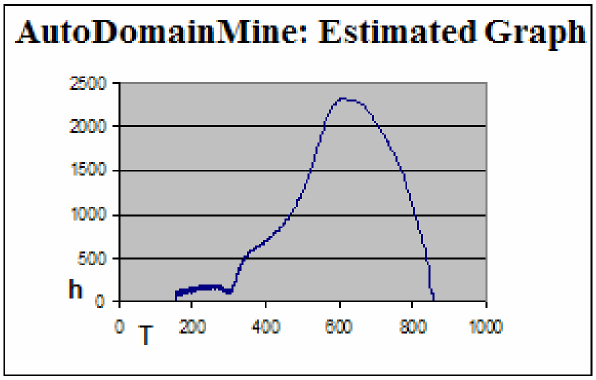

Figure 7: Output of Test 1 in Heuristic Approach

Figure 7 shows the estimated output of Test 1. On comparing this with the result of the real experiment, the users conclude that the estimation is close enough to the real result (see Figure 5). The maximum heat transfer occurs at approximately the same temperature and heat transfer coefficient values. Also the heat transfer curve in Figure 8 does not have a Leidenfrost point, nor does the curve from the real experiment. Most of the other regions on the two curves are also observed to be similar. Hence, this estimation is considered to be accurate.

Test 2: The input conditions in this test are the same as in Test 2 of the mathematical approach (see 6.1.1).

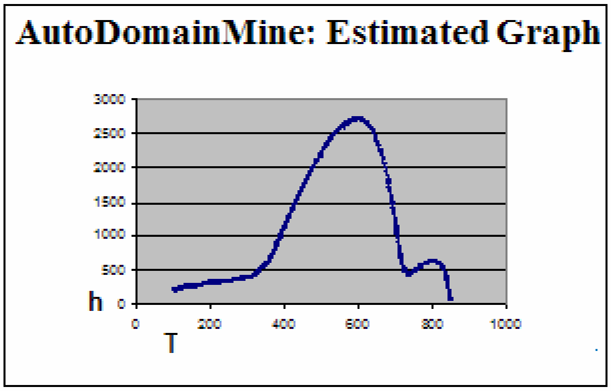

Figure 8: Output of Test 2 in Heuristic Approach

The output the estimation in Test 2 of the heuristic approach is shown in Figure 10. Upon comparing this with the result of the corresponding real experiment (see Figure 6), the users find that the two are close enough. The maximum heat transfer and Leidenfrost occur at around the same values of *h* and *T* for the real and estimated curves. Note that the portion of the curve on the left hand side of bell shaped-region is not considered significant as per the domain because the part has already been cooled to the quenchant temperature by then [16].

Thus, given that the significant points and most other regions are similar in the two curves this estimation is considered to be accurate.

Upon running tests with all the data in the test set, the estimation accuracy of AutoDomainMine is found to be in the range of 90 to 95%. Figure 9 shows the accuracy of this heuristic approach in the context of computational estimation and its targeted applications.

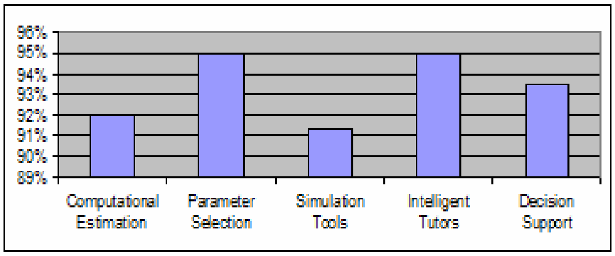

Figure 9: Accuracy of Heuristic Approach

Thus, on the whole, we find that the estimation accuracy in the heuristic approach is somewhat higher than that in the mathematical approach.

6.2. Efficiency

Efficiency refers to the amount of time taken to perform the estimation, i.e., the setup time for supplying the inputs and the response time of the tool. We record the amount of time taken to supply inputs for each test in both the mathematical and heuristic approaches. Note that in the mathematical approach, in addition to experimental input conditions, experts need to provide initial values for time-temperature. Thus, the time taken to provide these additional inputs is also recorded. The response time of each approach in terms of how long it takes to produce the output, given the inputs, is observed as well.

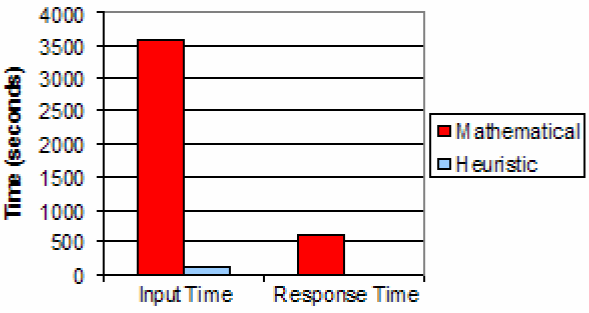

Figure 10: Efficiency of Estimation Approaches

Figure 10 shows average input and response times of the mathematical and heuristic approaches. We find that input time of the mathematical approach is much more than the heuristic approach. Also, the response time of the mathematical approach is of the order of minutes while that of the heuristic approach is negligible.

Thus, the heuristic approach is distinctly more efficient than the mathematical approach.

7. Related Work

An intuitive estimation approach is a naive similarity search over existing data [4]. Input conditions of a user-submitted experiment are compared with those of existing experiments to find the closest match. However, non-matching condition(s) could be significant in the domain. A weighted search [5] guided by basic domain knowledge can possibly be used to overcome this problem. Relative importance of search criteria, i.e., input conditions can be coded as weights into feature vectors. The closest match is then the weighted sum of matching conditions. However the problem is that these weights are not known apriori.

Case-based reasoning [1] could also be used for estimation. In our context, this involves comparing input conditions to retrieve the closest matching experiment, reusing its heat transfer curve as a possible estimate, performing adaptation if needed, and retaining the adapted case for future use. However adaptation approaches in the literature [1] are not feasible for us. For example in Heat Treating, if "agitation level" in the new case has a higher value than in the retrieved case, then a domain-specific adaptation rule could be used to infer that high agitation implies high heat transfer coefficients. However, this is not enough to plot a heat transfer curve in the new case.

Integration of rule-based and case-based approaches is another possible estimation approach [12]. However this is generally used when the case solution is categorical such as in medicine and law. To the best of our knowledge, it has not been used for graphs and images.

8. Conclusions

In this paper, mathematical and heuristic approaches for computational estimation are compared using the criteria of sample space, time complexity, data storage, efficiency and accuracy. We consider the Heat Treating domain and compare the direct-inverse heat conduction mathematical model with our proposed heuristic approach AutoDomainMine. Performance evaluation with real data from Materials Science is also presented. It is found that mathematical models are feasible when data from previous experiments is not stored, domain experts are available to provide inputs and efficiency is not critical. Heuristic approaches are found to give much higher efficiency and relatively higher accuracy than the mathematical models. However heuristic methods are applicable only when data from previously performed experiments is available. In the context of the estimation problem in this paper, heuristic approaches are preferred.

9. Acknowledgments

This work is supported by the Center for Heat Treating Excellence (CHTE) and by the Department of Energy - Award Number DE-FC-07-01ID14197.

10. References

[1] A. Aamodt and E. Plaza, "Case Based Reasoning: Foundational Issues, Methodological Variations & System Approaches", *Artificial Intelligence Communications*, 2003, Vol. 7, No. 1, pp. 39-59.

[2] J. V. Beck, B. Blackwell and C. R. St. Clair, *Inverse Heat Conduction*, Willey, NY, 1985.

[3] D. Bierman and P. Kamsteeg, *Elicitation of Knowledge with and for Intelligent Tutoring Systems,* Technical Report, University of Amsterdam, Amsterdam, Netherlands, 1988.

[4] J. Han and M. Kamber, *Data Mining: Concepts and Techniques,* Morgan Kaufmann, CA, 2001.

[5] D. Keim and B. Bustos, "Similarity Search in Multimedia Databases", *ICDE*, Boston, MA, Mar 2004.

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

[7] J. B. MacQueen, "Some Methods for Classification and Analysis of Multivariate Observations", *Mathematical Statistics and Probability*, Berkeley, CA, 1967, Vol. 1, pp. 281-297.

[8] J. R. Quinlan, "Induction of Decision Trees", *Machine Learning*, 1986, Vol. 1, pp. 81-106.

[9] S. Ma, M. Maniruzzaman and R. D. Sisson Jr., *Inverse Heat Conduction Problem in Estimating the surface Heat Transfer Coefficients by Steepest Descent Method,* Technical Report, Worcester Polytechnic Institute, Worcester, MA, Sep 2004.

[10] M. Maniruzzaman, A. S. Varde and R. D. Sisson Jr., "Estimation of Surface Heat Transfer Coefficients for Quenching Process Simulation", *MS&T*, OH, Oct 2006.

[11] A. Newell, J. C. Shaw and H. A. Simon, "Chess Playing Programs and the Problem of Complexity*, IBM Journal of Research and Development,* Vol. 4, No. 2, pp. 218-239.

[12] K. Pal and J. Campbell, "An Application of Rule-Based and Case-Based Reasoning in a Single Legal Knowledge-Based System", *Database for Advances in Information Systems,* 1997, Vol. 28, No. 4, pp. 48-63.

[13] R. K. Roy, *Design of Experiments Using The Taguchi Approach: 16 Steps to Product and Process Improvement*, Wiley, NY, Feb 2001.

[14] R. Ramakrishnan, J. Gehrke and V. Ganti, "Rainforest – A Framework for Fast Decision Tree Construction of Large Datasets", *Data Mining and Knowledge Discovery,* Vol. 4, pp. 127 – 162.

[15] S. Russell and P. Norvig, *Artificial Intelligence: A Modern Approach,* Prentice Hall, NJ, 1995.

[16] G. Stolz Jr., *Heat Transfer*, Wiley, USA, 1960.

[17] Scientific Forming Technologies Corporation. *DEFORM-HT*, Columbus, OH, USA, Feb 2005.

[18] R. Sisson Jr., M. Maniruzzaman and S. Ma, *Quenching: Understanding, Controlling and Optimizing the Process*, CHTE Seminar, Columbus, OH, Nov 2004.

[19] A. S. Varde, M. Takahashi, E. A. Rundensteiner, M. O. Ward, M. Maniruzzaman and Richard D. Sisson Jr., QuenchMiner™: *Decision Support for Optimization of Heat Treating Processes*, *IEEE IICAI,* Hyderabad, India, Dec 2003, pp. 993-1003.

[20] A. S. Varde, E. A. Rundensteiner, C. Ruiz, D. C. Brown, M. Maniruzzaman and R. D. Sisson Jr., "Integrating Clustering and Classification for Estimating Process Variables in Materials Science", *AAAI Poster Track*, Boston, MA, Jul 2006.

[21] A. S. Varde, M. Takahashi, E. A. Rundensteiner, M. O. Ward, M. Maniruzzaman and Richard D. Sisson Jr., "Apriori Algorithm and Game-of-Life for Predictive Analysis in Materials Science" *International Journal of Knowledge-Based & Intelligent Engineering Systems,* Vol. 8, No. 4, pp 213-228.

[22] A. S. Varde, *Graphical Data Mining for Computational Estimation in Materials Science Applications,* Ph.D. Dissertation, Worcester Polytechnic Institute, Worcestser, MA, Jun 2006.