Integrating Clustering and Classification for Estimating Process Variables in Materials Science

Aparna S. Varde[1,2], Elke A. Rundensteiner[1], Carolina Ruiz[1], David C. Brown[1,3], Mohammed Maniruzzaman[2,3] and Richard D. Sisson Jr.[2,3]

1. Department of Computer Science
2. Center for Heat Treating Excellence (CHTE)
3. Department of Mechanical Engineering
Worcester Polytechnic Institute (WPI), Worcester, MA 01609.
(508)-831-5681, aparna@wpi.edu, http://web.cs.wpi.edu/~aparna

Extended Abstract

The results of experiments in scientific domains such as Materials Science are often depicted as graphs. The *graphs* we refer to plot a dependent versus an independent variable showing the behavior of the experimental processes [5, 11]. They serve as good visual tools for analysis and comparison of the corresponding processes. Performing an experiment in a laboratory and plotting such graphs consumes significant time and resources motivating the need for computational estimation. This is precisely the aim of this research. More specifically, the research goals are as follows:

- Given the input conditions of an experimental process, estimate the resulting graph.
- Given the desired graph in an experimental process, estimate the input conditions to obtain it.

It is found that state-of-the-art approaches, e.g., case-based reasoning [2], mathematical modeling [5] and similarity search [1] do not give enough estimation accuracy in the targeted applications. The application domain of focus is the Heat Treating of Materials [5] that motivated this research. The graphs called heat transfer curves plot heat transfer (h) versus Temperature (T) during a rapid cooling process called quenching where heat transfer depicts the heat extraction capacity of a material [5].

We propose a computational estimation approach called AutoDomainMine [11] as depicted in Figure 1. The assumption is that existing experimental data is stored in a database as a set of input conditions and graph per experiment. AutoDomainMine integrates clustering and classification to discover knowledge from the data serving as the basis for estimation. Graphs from existing experiments are first clustered using a suitable clustering algorithm such as k-means [4]. Decision tree classification with algorithms such as ID3 / J4.8 [6] is then used to learn the clustering criteria (sets of input conditions characterizing each cluster) from which a representative pair of input conditions and graph is built per cluster. The decision trees and representative pairs form the knowledge discovered from existing experiments. Knowledge discovery is a one-time process. The discovered knowledge is used for the recurrent process of estimation. Given the input conditions of a new experiment, the relevant path of the decision tree is traced to estimate its cluster. The representative graph of that cluster is returned as the estimated graph for the experiment. Given a desired graph, the closest matching representative graph is found and its conditions are conveyed as estimated conditions to obtain the given graph. This estimation incorporates relative importance of conditions learned by decision trees. If a complete path does not match then a partial match is found based on higher levels of the tree using suitable thresholds.

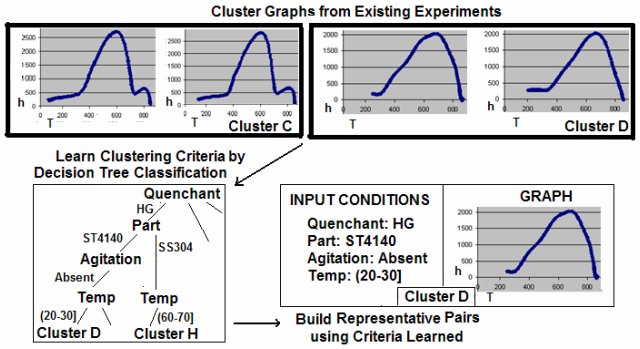

Figure 1: The AutoDomainMine Approach

AutoDomainMine follows a typical learning strategy of Materials Scientists. They often analyze by grouping experiments based on similarity of obtained graphs and reasoning causes of similarity group by group in terms of impact of input conditions on graphs [5, 8]. This learning strategy is automated for knowledge discovery in AutoDomainMine. The resulting estimation is found to be more accurate than state-of-the art methods [11].

A significant challenge in AutoDomainMine is capturing the semantics of the concerned graphs in clustering. Several distance metrics such as Euclidean and statistical distances exist in the literature [1]. However it is not known a priori which metric(s) would best preserve semantics if used as the notion of distance in clustering. Experts at best have vague notions about the relative importance of regions on the graphs but do not have a defined metric. State-of-the-art distance learning methods, e.g., [12] are either not applicable or not accurate enough in this context [10, 11]. We therefore propose a technique called LearnMet [10] (Figure 2) to learn semantics-preserving distance metrics for graphs. A LearnMet metric *D* is a weighted sum of components where each component is an individual metric such as Euclidean or statistical distance (or a domain-specific metric [10]), and its

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 # DE-FC-07-01ID14197.

weight gives its relative importance in the domain. LearnMet iteratively compares a training set of actual clusters given by experts with predicted clusters obtained from any fixed clustering algorithm, e.g., k-means [4]. In the first iteration, a guessed metric *D* is used for clustering. This metric is adjusted based on error between predicted and actual clusters using our proposed Weight Adjustment Heuristic [10] until error is below a given threshold or a maximum number of epochs is reached. The metric with error below threshold or with minimum error among all epochs is returned as the learned metric. The output of LearnMet is used as the notion of distance for clustering the graphs.

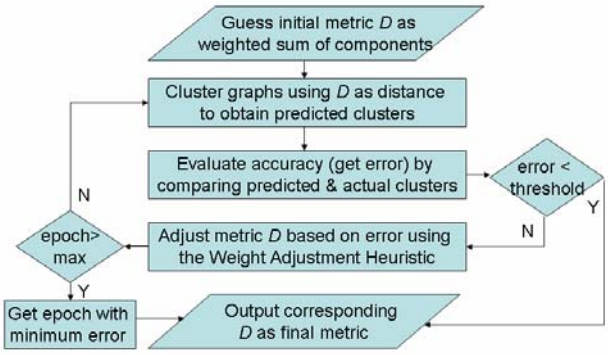

Figure 2: The LearnMet Technique

Another challenge in AutoDomainMine is capturing relevant data in each cluster while building representatives. A default approach of randomly selecting a representative pair of input conditions and graph per cluster is not found to be effective in preserving the necessary information. Since several combinations of conditions lead to a single cluster, randomly selecting any one as a representative causes information loss. Randomly selected representatives of graphs do not incorporate semantics and ease of interpretation based on user interests. Thus, we propose a methodology called DesRept [11] to design domain-specific cluster representatives. In DesRept, two design methods of guided selection and construction are used to build candidates such as medoid and combined representatives [11]. Candidates are compared using our proposed DesRept Encoding analogous to the Minimum Description Length principle [7]. The criteria in this encoding are complexity of the representative and information loss due to it based on user interests. Winning candidate(s) with the lowest encoding is/are output as designed representative(s).

Quenchant	Part	Agitation	Temp	Oxide
DHR88A	SS304	Absent	(140 - 150]	None
DHR88A	ST4140	Absent	(200 - 210]	Thin
DW25	ST4140	Absent	(190 - 200]	Thin
HG	ST4140	Absent	(150 - 160]	Thin
Mar-355	ST4140	Absent	(140 - 150]	Thin
T7A	ST4140	Absent	(170 - 180]	Thin

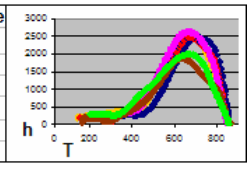

Figure 3: An Example of the Output of DesRept

An example of a designed representative pair output from DesRept is shown in Figure 3. This is called a combined representative [11]. It depicts all sets of conditions leading to the cluster sorted using domain knowledge along with a superimposed graph of all graphs in the cluster. Likewise, various designed representative pairs showing information in different levels of detail are output by DesRept. These are used as the basis for estimation in AutoDomainMine [11].

Thus the main contributions of this research are:

- The basic AutoDomainMine estimation approach of integrating clustering and classification thereby automating a learning strategy of scientists.
- The LearnMet technique for learning semantics-preserving distance metrics for graphs, in particular its Weight Adjustment Heuristic
- The DesRept methodology for designing domain-specific cluster representative pairs along with the DesRept Encoding for evaluating them.

AutoDomainMine has been evaluated rigorously in the Heat Treating domain. AutoDomainMine estimation is compared with real experiments from a distinct set not used for training. If the real data matches the estimation within a given threshold (10% here) then the estimation is considered to be accurate. Accuracy is reported as the percentage of accurate estimations over the test set [11]. It is observed that the basic AutoDomainMine approach [8] of integrating clustering and classification gives an average estimation accuracy of 75% which is higher than state-of-the-art methods such as similarity search (65% at best). The estimation accuracy increases to around 86% when metrics learned from LearnMet [10] are used for clustering. Cluster representatives designed by DesRept [11] further enhance estimation accuracy to around 91%. Different designed representatives are useful in various applications, e.g., decision support [9] and simulation [3].

References

[1] Keim, D. and Bustos, B.: Similarity Search in Multimedia Databases. *IEEE's ICDE,* 2004, pp. 873-874.

[2] Koldner, J.: *Case-Based Reasonin*g. Morgan Kaufmann, 1993.

[3] 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.

[4] MacQueen J.: Some Methods for Classification and Analysis of Multivariate Observations. *Mathematical Statistics and Probability,* 1967, Vol. 1, pp. 281 – 297.

[5] Mills, A.: *Heat and Mass Transfer.* Richard Irwin, 1995.

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

[7] Rissanen, J.: Stochastic Complexity and the MDL Principle. *Econometric Reviews,* 1987, Vol. 6, pp. 85-102.

[8] Sisson, R., Maniruzzaman, M., Ma, S., and Varde, A.: *Quenching: Understanding, Controlling and Optimizing the Process.* Technical Report: 2004, TR# CHTE-II-04.

[9] Varde, A., Takahashi, M., Rundensteiner, E., Ward, M., Maniruzzaman, M. and Sisson, R.: QuenchMiner™: Decision Support for Optimization of Heat Treating Processes. *IEEE's IICAI,* 2003, pp. 993 - 1003.

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

[11] Varde, A.: *Graphical Data Mining for Computational Estimation in Materials Science Applications.* Ph.D. Dissertation in progress, WPI, Worcester, MA, 2006.

[12] Xing, E., Ng. A., Jordan, M. and Russell S.: Distance Metric Learning with Application to Clustering with Side Information. *NIPS, 2003,* pp. 505-512.