Using Uncertain Chemical and Thermal Data to Predict Product Quality in a Casting Process

Catarina Dudas
Virtual Systems Research Centre
Box 408, 541 28 Skövde
Sweden
+46 500 44 85 83
catarina.dudas@his.se

Henrik Boström
Informatics Research Centre
Box 408, 541 28 Skövde
Sweden
+46 70 523 85 84
henrik.bostrom@his.se

ABSTRACT

Process and casting data from different sources have been collected and merged for the purpose of predicting, and determining what factors affect, the quality of cast products in a foundry. One problem is that the measurements cannot be directly aligned, since they are collected at different points in time, and instead they have to be approximated for specific time points, hence introducing uncertainty. An approach for addressing this problem is investigated, where uncertain numeric feature values are represented by intervals and random forests are extended to handle such intervals. A preliminary experiment shows that the suggested way of forming the intervals, together with the extension of random forests, results in higher predictive performance compared to using single (expected) values for the uncertain features together with standard random forests.

1. INTRODUCTION

Data mining techniques have become standard tools to develop predictive and descriptive models in situations where one wants to exploit data collected from earlier observations in order to optimize future decision making [8]. One of the key characteristics of data mining is that it can be used for analyzing data that has been collected during the normal operations of a process, i.e., data does not have to be specifically collected for this purpose [8].

However, one drawback of having collected data without analysis in mind is that the data may not be optimal for the intended purpose. The collected data can for instance lack measurements on individual object level, i.e., information is instead represented on batch level. This is typical for the application area considered in this study, which has been done in cooperation with Volvo Powertrain in Skövde, Sweden, a supplier of power train parts, such as cylinder blocks, gear boxes and drive shafts, to the business areas within the Volvo Group. In this study, we focus on the process line of cylinder heads that are casted in the foundry. In particular, we study how to derive a classification model for the quality of cast cylinder heads by analyzing process and casting data.

Not only is the classification model of interest for making predictions, but also for identifying variables affecting the quality. In this application, different types of data are stored in different databases, and one major difficulty is to merge process data, which is on individual product level, with casting data, which is on a batch level, hence introducing uncertainty regarding measurements for the individual products.

In the next section, we describe the application in more detail, including the data sources considered. In section three, we point out sources of uncertainty in the merged data and discuss approaches to dealing with these. In section four, we present a preliminary experiment that has been conducted. Finally, in section five, we give some concluding remarks and point out directions for future research.

2. CASE STUDY DESCRIPTION

The considered cylinder head process line at Volvo Powertrain consists of three sub-lines with two different marriage points. A marriage point is where two sub-lines intersect and parts from each line are assembled.

2.1 Explanation of the Casting Process Line

The casting process can be divided into five main actions; pattern making, core making, molding, melting and cleaning.

The pattern is a model of the cylinder head which is used for the form shooting. The form is produced by packing molding sand under great pressure and the result is an inverse image of the final product. It consists of one collecting form (bottom half), where the cores are placed, and one covering form (top half).

A cylinder head is not a solid product; i.e., it has interior cavity. To achieve the cavities in the cylinder head, inner parts, called cores, must be added to the form. After the shooting of the cores, these are glued together with the collecting form at the first marriage point. The void space between the form and the cores is what the final product develops into.

The assembling of the cores and the form is part of the molding process. This involves preparation of the forms in order for these to get ready for the melted material. The last step in the molding process is to put the covering form on top of the collecting form before the melt is poured into the assembled final form. This is the second marriage point in the cylinder head process line.

The melting process takes places at another area in the foundry, since it supplies several different lines with melted metal, not only the cylinder head line. The melted metal is prepared during the melting process to obtain the appropriate characteristics of the

U'09, June 28, 2009, Paris, France.

melt. Such preparations involve temperature regulations and addition of the proper amount of chemicals. The melted metal is then transported to the casting area where the metal is poured into the form. The casting area can hold 24 forms and these are divided into 12 chills where each chill has two bins. A chill holds the form when the melt is poured into it and supports the solidification of the melted metal.

Cleaning is the final step in the casting process and is done when the melted material has been cooled down sufficiently enough. This step refers to sand removal; the casted product is separated from the form and the sand is removed. If the product passes a quality control, it is transported to the next line where the cylinder head is welded; i.e., to improve the surface of the cylinder head.

2.2 Complexity of the Casting Process

The casting process is complex with multivariate interactions of known but also unidentified factors which makes it practically impossible for humans to grasp, as it has been observed that humans are normally not able to simultaneously analyze situations involving more than three variables effectively and this becomes even more difficult when the data are corrupted by noise and uncertainty [7]. The current way of analyzing the casting process is done in a one-variable-at-a-time manner. Due to the complex relationships, there is a requirement for a more sophisticated way of analyzing data from such processes and it is believed that data mining can achieve this in a useful way.

2.3 Data Used in an Initial Experiment

In a first experiment, process data and quality data was collected from the processing line before casting of cylinder heads, see Figure 1. The process and quality data is stored on individual product level, i.e., there are corresponding process values for each cast cylinder head and its final quality outcome. The process data consists of about 100 variables, such as pressure, time taken between two process steps, weight, what casting chill was used and used fixture, i.e., a tool which holds the form during the process line.

The casted product fails the quality inspection if some fault can be found on the product. Each rejected product is given one of about 50 rejection codes, but in this study all rejection codes are treated identically; i.e., the quality data is transformed into binary data (*discard* or *no-discard*).

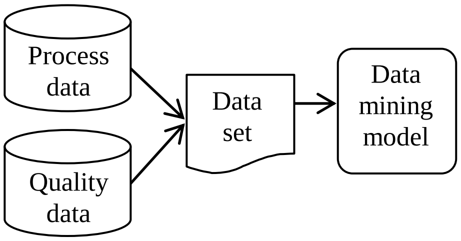

Figure 1. The data mining process in the initial study.

A predictive data mining model was built using the Rule Discovery System (RDS) [6]. In addition to generating predictive models, e.g., random forests, this software provides some insight into what factors are of importance by presenting the variable importance of each independent variable, i.e., how much the variable, relative to all other variables, contributes to reducing the error of the dependent variable.

Some of the process variables or settings were strongly believed to have some impact on the quality of the resulting product, i.e., the cylinder head, but this could not be confirmed by the analysis. It was concluded that the use of process data only is not sufficient to get a good and accurate model. It was decided that the process dataset should be extended with chemical and thermal analysis data.

2.4 Casting Analysis Data

In retrospect, one may consider the idea of predicting the quality based only on data originating from the pre-processes of the mold before the casting to be rather naïve. The casting process is very complex, and the melt itself can be expected to have an even greater impact on the quality of the cylinder heads than the subsequent process. As a next step, it is inevitable to also include data from the casting process, as depicted in Figure 2.

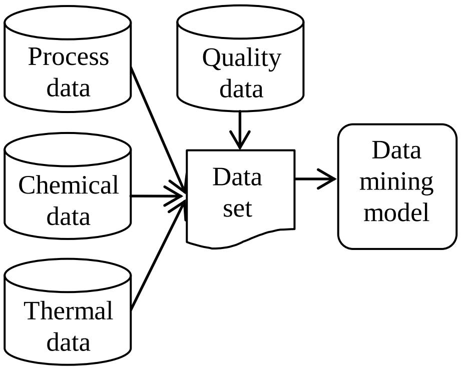

Figure 2. Extended data mining process.

The casting process starts with melting down raw material in a melting oven. The melt is then transported to a buffer oven, where the composition of the melt is controlled; the melted material is prepared by e.g., regulating the temperature and adding chemicals. The melt is transferred to a final oven, the casting oven, and a casting analysis is conducted when this oven is refilled with a new melt. One refill of this oven contains melt sufficient for casting of eight cylinder heads.

2.4.1 Chemical analysis data

Analysis of the chemical compound of a melt is undertaken at every second or third refill of the casting oven. This would imply that one chemical analysis can be linked to 16-24 cylinder heads. For a cylinder head that is cast in a near time to the analysis, this measurement will most likely be close to what would be obtained at time of casting, but the deviations are most likely larger for cylinder heads that are cast later.

The chemical analysis measures the substance levels of 18 different chemicals, including carbon, manganese, sulfur and nickel. According to human expertise, the chemical compound of the melt is expected to have a great impact on the final product quality, but the relationships between substance levels and quality of casting are far from clear.

2.4.2 Thermal analysis data

A sample of molten material is collected and analyzed from a material structure point of view. The thermal analysis data contains measurements of the cooling of the molten material. The various events that arise during the cooling process are considered to play a major role in the quality of the final cylinder head. The cooling curve represents the change in temperature over time and its first and second derivatives can also be used to gain information of the thermal conditions in the melt.

One such measurement is the *recalescence*, which is a temporary increase in heat during the cooling process and this may provide a lot of information about the inoculation of the iron. The inoculation of the molten material is used in order to get a desired crystal structure, and is accomplished by adding substances in the melt to ease the manifestation of a uniform crystal structure.

Thermal analysis is supposed to be conducted each time the casting oven is refilled and therefore 8 cylinder heads can be connected to each thermal analysis.

3. UNCERTAINTY IN THE DATA

In many real-world datasets, the problem of handling uncertain data arises, such as missing values, noise, etc. In this section, we first show how combining data from different data sources can further promote uncertainty, e.g., as discussed in [3], and we then present different approaches for addressing this problem.

3.1 Merging Process and Casting Data

The process data contains some missing data, but this is not of great concern in this study, since RDS, similar to many other systems, including e.g. C4.5 [5], can directly handle this type of data. All data in the process database can furthermore be connected to one specific cylinder head.

As mentioned in section 2.4.1 and 2.4.2, the casting analyses are planned for being taken at regular occasions, but in reality they are less frequent. One chemical analysis can therefore be linked to up to approximately 30 cylinder heads and the number of cylinder heads connected to each thermal analysis is approximately 19 on average. An illustration of this is shown in Table 1, where the first and second columns correspond to a product ID of the cylinder head and time of casting, respectively. To each of these product IDs, there are about 100 associated process parameters, as described in section 2.3. The last two columns show time points when chemical and thermal data is collected, and as can be seen, there is a lack of measurements between consecutive analyses of chemical and thermal data.

3.2 Different Approaches

The process data can be used with no further preparation, but the missing chemical and thermal data need to be taken special care of. In addition to the straightforward, but not very productive, alternative of treating these data as missing, which hardly would result in that any chemical and thermal data for the cylinder heads are provided, there are a number of additional strategies for handling the uncertain data, which are described in this section.

Table 1. A sample of the frequency of the casting analysis.

Product ID	Process data	Chemical data	Thermal data
84001	19:40		
84002	19:52	19:43	
84003	19:53		
84004	19:57		
84005	19:58		
84006	20:02		
84007	20:02		
84008	20:07		
84009	20:08		
84010	20:12		
84011	20:13		
84012	20:18		
84013	20:19		
84014	21:04		20:43
84015	21:04		
84016	21:09		
84017	21:10		
84018	21:14		
84019	21:15		
84020	21:19	21:18	
84021	21:19		
84022	21:23		
84023	21:24		
84024	21:29		21:26
84025	21:30		

3.2.1 Use identical data for several products

One way of handling the missing data is to use the last measurements for all subsequent products until a new analysis is conducted.

This will result in a data set that typically has unique values for the process data, but in which the chemical and thermal data will be identical for several products. The benefit of this method is that the standard data mining methods can be applied directly on the resulting dataset. The disadvantage is that a product cast shortly after the analysis will have identical values to products cast much later. For example; product 84002 and 84019 in Table 1 will have identical chemical values, but the time span between them is 83 minutes, and the chemical substances have most likely changed during this time period, in particular the carbon content.

3.2.2 Interval-based data

Another approach would be to use intervals representing the uncertainty of a measurement, i.e., the true value is expected to reside within the interval. When forming intervals around the most recent chemical or thermal estimate, it appears natural that these should be narrower for products that have been casted close in time to the measurement, compared to products that have been casted long after, i.e., the uncertainty is expected to increase with time. When extending the process measurements, which are represented by exact numbers, with chemical and thermal data represented by intervals, two main issues arise. The first concerns how to form the intervals. There are several possible alternatives for this. One approach is to try to model the error of estimated parameters as a function of time. This requires that a suitable model class is chosen, e.g., linear function, and that its parameters can be determined using available training data.

The second issue concerns the inability of most current data mining systems to handle interval data. For this study, the Rule Discovery System (RDS) [6] has been extended to handle intervals for numeric features, in the following way.

Each numeric feature value for an example is represented by a pair *Value*/*Error*, where *Value* is the expected (most likely) value of the feature for the example according to some measurement, and *Error* denotes the size of the interval surrounding the expected value, where *Lower*=*Value*-*Error*/2 and *Upper*=*Value*+*Error*/2 are the lower and upper bounds of the value, i.e., the true value is expected (to some suitable degree of confidence) to appear in this interval.

The Rule Discovery System may generate decision trees [5], as well as ensembles of such trees (or random forests [1]), from both numeric and nominal features, for both classification and regression. Missing values are handled by distributing fractional examples over multiple nodes, as originally suggested in [5], with the weight of each fraction corresponding to the relative frequency of examples with a known particular value. Standard numeric features are discretized during tree growth, as proposed in [2].

However, instead of just using the expected value for choosing a single child node to place an example in when growing a tree, three approaches to utilizing the intervals are considered.

The first two approaches distribute fractions of an example over multiple child nodes, similar to when having missing values, but where the weight of each fraction is determined in the following way. For a split of a node into two children *Left* and *Right* using the conditions *Variable≤Threshold* and *Variable>Threshold* respectively, and an example with associated values *Lower* and *Upper*, the fraction of the example going into *Left* is one if *Upper≤Threshold*, while the fraction of the example going into *Right* is one if *Lower>Threshold*, and otherwise the fraction of the example going into *Left* is *F* = (*Threshold*-*Lower*)/(*Upper*-*Lower*) and the fraction going into *Right* is 1-*F*. This assumes that the probability of the true value is uniformly distributed over the entire interval. For example, if the conditions *V≤12* and *V>12* are associated with the left and right child respectively, then an example having the value *V=14* with error *8*, will be distributed such that (12-10)/(18-10) = 0.25 of the weight falls into the left child and 0.75 of the weight falls into the right. The difference between the two first approaches concerns the forming of thresholds for conditions involving numerical features, i.e., *Variable≤Threshold* or *Variable>Threshold*. At each node, when a numeric feature is to be discretized according to the method proposed in [2], the first approach considers only the expected value, i.e., *Value* above, of each training example. The second approach instead samples a value from the interval for each training example. For example, assume we are given the following sequence of triples, where each triple *(V,E,C)* represents an example having the value *V* with error *E* on a feature and class label *C*:

(6,4,+), (8,6,+), (10,4,-), (12,4,-)

The first approach, which only considers the expected values, will in this case only result in one possible threshold, i.e., 9, separating differently labeled examples according to the method in [2]. The second approach, which randomly assigns values from the intervals, may in this case not only find a single threshold, but several. For example, if the random assignments from the above intervals are 7, 11, 9 and 13 respectively, then there are three thresholds separating differently labeled examples: 8, 10 and 12.

The third approach to handling the intervals does not consider distributing fractions of examples over multiple nodes, other than for completely missing feature values, but instead randomly selects a value from each interval for each example, prior to growing a tree. Hence, during tree construction, no particular care is taken to uncertain feature values, and the standard discretization technique is employed. When generating a forest of trees, a new assignment of values from the intervals is made for each tree in the forest.

4. PRELIMINARY EXPERIMENT

We first describe the experimental setup and then present results from a comparison of the approaches for handling uncertain data as described in the previous section.

4.1 Experimental Setup

A dataset consisting of 13497 examples was assembled, where each example (cylinder head) was represented using 60 features, of which 19 correspond to process parameters, 22 concern chemical analysis measurements and 19 concern thermal analysis measurements.

Each interval of a numeric feature value was represented by a term *Value*/*Error*, using the most recent measurement value as *Value*, while *Error* was computed as a function of time between the most recent measure and the casting time t_i for product *i* and the standard deviation *sd* for each thermal and chemical variable, as shown in eq. 1.

$$Error = \frac{1}{1+e^{a \cdot t_i + b}} \cdot sd \qquad (1)$$

The constants *a* and *b* in the error function were chosen so that $e^{a \cdot ti+b}$ becomes very large for t_i close to zero and is close to 0 for the maximum value of t_i. To achieve this, the value of *b* should be sufficiently large when t_i is close to zero and $a \cdot t_i+b$ small enough (less than zero) for large t_i. In this experiment, the constants were chosen to be *a = -1* and *b = 8*. By using this formula, the length of the interval becomes narrower for products that have been cast more recent to the measuring time, while the width increases for products cast longer after. *Error* is close to zero for products cast close to the measurement and up to one standard deviation for products that have been cast long after.

The classification task in the experiment is binary with the two classes *discard* and *no-discard*. The class frequencies are highly imbalanced, with less than 5% belonging to the former class, making it difficult to obtain a higher accuracy than the default classifier, i.e., classifying everything as *no-discard*. We hence decided to use as evaluation criterion the area under the ROC curve (AUC), i.e., the probability that an example belonging to a class is ranked as being more likely belonging to the class than an example not belonging to the class [4].

In the experiment, stratified 10-fold cross validation was employed to estimate the AUC. Four versions of a random forest with 100 trees was generated with the Rule Discovery System: the first ignoring the intervals and only considering the expected values, hence effectively implementing the first strategy described in section 3.2.1, while the three other versions employ the three different approaches to utilize intervals as described in section 3.2.2. Exactly the same training and test data, as well as random seeds for (bootstrap) selection of examples and features were used for all versions.

4.2 Experimental Results

The approach that considers only the expected values for the uncertain features obtained an AUC of 62.1. The three approaches that utilize the intervals obtained a higher AUC, ranging from 62.8 for the first of these (distributing fractions of examples and using expected values to form thresholds), through 63.0 for the second (distributing fractions of examples and using random assignments to form thresholds) to 63.9 for the third (random assignments prior to tree growth). This shows that some performance improvement may indeed be obtained by the suggested way of representing uncertain feature values by intervals, and approximating the error as a function of time since measurement, compared to using only the expected value.

5. CONCLUDING REMARKS

We have investigated approaches for merging uncertain chemical and thermal analysis data with process data in order to obtain predictive models for quality of cast cylinder heads. A straightforward approach of representing uncertain feature values by intervals, and approximating the error as an exponential function of time that has passed since the measurement, was shown to give some performance improvement with respect to area under ROC curve, when used together with random forests extended with the ability to handle interval-based data, compared to using the last measurement only of the uncertain features.

Three different ways of extending random forests to utilize intervals for uncertain numeric features were investigated. For the current dataset, it turned out to be beneficial to randomly assign values from the intervals before growing each tree in a forest, rather than distributing fractions of examples during tree growth. Future work includes investigating whether this finding also holds for other datasets. Another direction for future work is to relax the assumption that the probability of the true feature value is uniformly distributed within an interval, and extend the methods to handle also non-uniform, e.g., normal, distributions.

The result of using only one quite straightforward approach of approximating the size of the error has been presented here. There are obviously many other potential ways of doing this. In particular can the choice of the constants *a* and *b* in eq. 1 be altered so that time will have a somewhat different impact on the observations. Keeping the scaling factor in the error function within the range of [0, 1] will not affect the maximum interval length. Hence, changing the number of standard deviations in the error function is yet another option, although it will not change the interrelationship among the observations, but only the interval length.

Besides the alternative choices of the parameters *a*, *b* and number of standard deviations, future work also includes investigating how to select a proper error function for each uncertain variable. Different error functions and parameter settings might be appropriate for different chemical and thermal variables.

ACKNOWLEDGMENTS

This work was supported by the Information Fusion Research Program (www.infofusion.se) at the University of Skövde, Sweden, in partnership with the Swedish Knowledge Foundation under grant 2003/0104.

REFERENCES

[1] Breiman L., (2001). Random Forests, Machine Learning, Vol. 45, Issue 1, 5-32

[2] Fayyad U. and Irani K. 1992. On the Handling of Continuous-Valued Attributes in Decision Tree Generation. Machine Learning 8: 87-102

[3] Graham Cormode and Andrew McGregor. 2008. Approximation Algorithms for Clustering Uncertain Data, Symposium on Principles of Database Systems 2008.

[4] Provost F., Fawcett T. and Kohavi R. 1998 The case against accuracy estimation for comparing induction algorithms, Proc. Fifteenth Intl. Conf. Machine Learning, 445-553

[5] Quinlan J.R. 1993. C4.5: Programs for Machine Learning, Morgan Kauffman, San Francisco

[6] Rule Discovery System, v. 2.6.0, Compumine AB, http://www.compumine.com/web/public/rds (accessed April 17, 2009)

[7] Wang, X. Z. 1999. Data Mining and Knowledge Discovery for Process Monitoring and Control, Springer-Verlag, London.

[8] Witten I. and Frank E. 2005. Data Mining: Practical Machine Learning Tools and Techniques (Second Edition), Morgan Kaufmann Publisher, San Francisco.