Data Mining for Improving Manufacturing Processes

Lior Rokach
Ben-Gurion University, Israel

INTRODUCTION

In many modern manufacturing plants, data that characterize the manufacturing process are electronically collected and stored in the organization's databases. Thus, data mining tools can be used for automatically discovering interesting and useful patterns in the manufacturing processes. These patterns can be subsequently exploited to enhance the whole manufacturing process in such areas as defect prevention and detection, reducing flow-time, increasing safety, etc.

When data mining is directed towards improving manufacturing process, there are certain distinctions that should be noted compared to the classical methods employed in quality engineering, such as the experimental design. In data mining the primary purpose of the targeted database is not data analysis; the volume of the collected data makes it impractical to explore it using standard statistical procedures (Braha and Shmilovici, 2003).

BACKGROUND

This chapter focuses on mining performance-related data in manufacturing. The performance can be measured in many different ways, most commonly as a quality measure. A product is considered as faulty when it does not meet its specifications. Faults may come from sources such as, raw material, machines setup and many other sources.

The quality measure can either have nominal values (such as "good"/"bad") or continuously numeric values (Such as the number of good chips obtained from silicon wafer or the pH level in a cream cheese). Even if the measure is numeric, it can still be reduced to a sufficiently discrete set of interesting ranges. Thus we can use classification methods in order to find the relation between the quality measure (target attribute) and the input attributes (the manufacturing process data).

Classification methods can be used to improve the learning curve both in the learning pace, as well as in the target measure that is reached at the mature stage. The idea is to find a classifier that is capable of predicting the measure value of a certain product or batch, based on its manufacturing parameters. Subsequently, the classifier can be used to set up the most appropriate parameters or to identify the reasons for bad measures values.

The manufacturing parameters obviously include the characteristics of the production line (such as which machine has been used in each step, how each machine has been setup, operation sequence etc.), and other parameters (if available) relating to the raw material that is used in the process; the environment (moistness, temperature, etc); the human resources that operate the production line (the experience level of the worker which have been assigned on each machine in the line, the shift number) and other such significant factors.

The performance measure (target attribute) in manufacturing data tends to have imbalanced distribution. For instance, if the quality measure is examined, then most of the batches pass the quality assurance examinations and only a few are considered invalid. On the other hand, the quality engineer is more interested in identifying the invalid cases (the less frequent class).

Traditionally, the objective of the classification method is to minimize the misclassification rate. However, for the unbalanced class distribution, accuracy is not an appropriate metric. A classifier working on a population where one class ("bad") represents only 1% of the examples can achieve a significantly high accuracy of 99% by just predicting all the examples to be of the prevalent class ("good"). Thus, the goal is to identify as many examples of the "bad" class as possible (high recall) with as little false alarms (high precision). Traditional methods fail to obtain high values of recall and precision for the less frequent classes, as they are oriented toward finding global high accuracy.

Figure 1. Decision tree for quality assurance

```
CW_WASH_DUR <= 286
|  FINAL_COOLING_TEMP <= 5.9
|  |  AVG_COOLING_TEMP <= 10.1: Tasty (0.864,0.136)
|  |  AVG_COOLING_TEMP > 10.1: Sour (0.323,0.674)
|  FINAL_COOLING_TEMP > 5.9
|  |  AVG_COOLING_TEMP <= 12.3: Tasty (0.682,0.318)
|  |  AVG_COOLING_TEMP > 12.3: Sour (0.286,0.714)
CW_WASH_DUR > 286: Tasty (0.906,0.094)
```

Usually in manufacturing plants there are many input attributes that may affect performance measure and the required number of labelled instances for supervised classification increases as a function of dimensionality. In quality engineering mining problems, we would like to understand the quality patterns as soon as possible in order improve the learning curve. Thus, the training set is usually too small relative to the number of input features.

MAIN FOCUS

Classification Methods

Classification methods are frequently used in mining of manufacturing datasets. Classifiers can be used to control the manufacturing process in order to deliver high quality products.

Kusiak (2002) suggested a meta-controller seamlessly developed using neural network in order to control the manufacturing in the metal industries. While neural networks provide high accuracy, it is usually hard to interpret its predictions. Rokach and Maimon (2006) used a decision tree inducer in cheese manufacturing. As in every dairy product, there is a chance that a specific batch will be found sour when consumed by the customer, prior to the end of the product's shelf-life. During its shelf-life, the product's pH value normally drops. When it reaches a certain value, the consumer reacts to it as a spoiled product. The dairy department performs randomly tests for pH as well organoleptic (taste) at the end of the shelf-life.

Figure 1 demonstrates a typical decision tree induced from the manufacturing database. Each representing symptoms of a product quality, denoted here as features Cold Water Wash Duration, Final Cooling Temperature and Average Cooling Temperature. The leaves are labeled with the most frequent class together with their appropriate probability. For instance, the probability to get a tasty cheese is 0.906 if the Cold Water Wash Duration is greater than 286 seconds.

Ensemble of Classifiers

The main idea of ensemble methodology is to combine a set of models, each of which solves the same original task, in order to obtain a better composite global model, with more accurate estimates than can be obtained from using a single model.

Maimon and Rokach (2004) showed that ensemble of decision trees can be used in manufacturing datasets and significantly improve the accuracy. Similar results have been obtained by Braha and Shmilovici (2002) arguing that ensemble methodology is particularly important for semiconductor manufacturing environments where various physical and chemical parameters that affect the process exhibit highly complex interactions, and data is scarce and costly for emerging new technologies.

Because in many classification applications non-uniform misclassification costs are the governing rule, has led to a resurgence of interest in cost-sensitive classification. Braha et al. (in press) present a decision-theoretic classification framework that is based on a model for evaluating classifiers in terms of their value