Optimising operational costs using Soft Computing techniques

Javier Sedano\textsuperscript{a}, Alba Berzosa\textsuperscript{a}, José R. Villar\textsuperscript{b*}, Emilio Corchado\textsuperscript{c}, Enrique de la Cal\textsuperscript{b}

\textsuperscript{a}Grupo de Investigación de Inteligencia Artificial y Electrónica Aplicada, Instituto Tecnológico de Castilla y León, Polígono Industrial de Villalonquéjar C/López Bravo, 70, 09001, Burgos, Spain
\textsuperscript{b}Departamento de Informática, Universidad de Oviedo, Campus de Viesques s/n, 33204, Gijón, Spain
\textsuperscript{c}Departamento de Informática y Automática, Universidad de Salamanca, Plaza de la Merced s/n, 37008, Salamanca, Spain

Abstract. A Manufacturing Execution System (MES) consists of high-cost, large-scale, multi-task software systems. Companies and factories apply these complex applications for the purposes of production management to monitor and track all aspects of factory-based manufacturing processes. Nevertheless, companies seek to control the production process with even greater rigour. Improvements associated with an MES involve the identification of new knowledge within the data set and its integration in the system, which implies a step forward to Business Process Management (BPM) systems, from which the users of an MES may gain relevant information, not only on execution procedures but to decide on the best scheduled arrangements. This work studies the data gathered from a real MES that is used in a plastic products factory. Several Artificial Intelligence and Soft Computing modelling methods based on fuzzy rules assist in the calculation of manufacturing costs and decisions over shift work rotas: two decisions that are of relevance for the improvement of the execution system. The results of the study, which identify the most suitable models to facilitate execution-related decision-making, are presented and discussed.

Keywords: Applied Soft Computing, Artificial Intelligence, Enterprise Resource Planning, Manufacturing Execution Systems

1. Introduction

Enterprise Resource Planning (ERP) and Manufacturing Resource Planning (MRP) [74, 113] are well-known software applications currently used in production and cost optimization in factory plants. These platforms are based on the available Information Technology (IT) frameworks [61, 62]. The increase of the IT presence in industry has resulted in the growth of the aforementioned applications and their integration in production systems [82, 89, 94, 108, 119].

Manufacturing Execution Systems (MES) are IT systems that are used for management resource planning: equipment, employees and inventories [18, 112].

An MES may be implemented in the context of a production control system or a manufacturing monitoring and supervisory system. In the former case, the objective is to provide the company with a research laboratory for products and processes, whereas in the latter case, the MES is a computer-aided system that assists with decision-making processes that relate to manufacturing.

However, designing and deploying a user-friendly MES which fulfils the above-mentioned objectives represents a significant challenge, owing to a large extent to the complexity of modern production systems, plants and products [28, 90].

Soft computing [6, 19, 21, 32, 34, 61, 80, 111] is a collection or set of computational techniques in machine learning [1, 4, 42, 86], such as artificial neural networks [3, 7, 8, 11, 12, 13, 15, 40, 43, 47, 48, 49, 31, 55, 56, 57, 64, 77, 78, 79, 83, 101, 104, 109], genetic algorithms [9, 10, 16, 22, 23, 26, 29, 60, 67, 70, 72, 73, 95, 96, 97, 98], fuzzy systems [5, 6, 20, 54, 58, 68, 71, 87, 100, 121, 114], simulated annealing algorithms [122], spiking neural networks [2, 41,
59, 76, 91, 99, 106, 107], case-based reasoning [65, 105, 116], and swarm intelligence [35, 36, 37, 88, 115, 118], which investigate, simulate and analyze very complex issues and phenomena concerned with problems in which tractability, robustness and uncertainty are significant factors [46]. There is also a significant number of recent articles on developing hybrid systems through the integration of various soft computing techniques [44, 50, 63, 81, 110, 117] or integration with other techniques such as wavelets [14, 39, 92].

This study applies several soft computing techniques and analyses its results. Its aim is to obtain classification models that, according to different manufacturing conditions –machines, products, stoppage time, run time, micro stoppage time...-, support staff at a plastic products factory with budgeting – costs management- and operator shift rota. Nevertheless, the final and mid-term aim of this study is to extend its scope and integrate the models into the company’s MES, in order to enhance the capability to estimate real operational costs under the best conditions for execution. Section 2 describes the problem scenario, and subsequent subsections consider the objective of the study and similar approaches in the literature. Section 3 explains the fundamentals of the different methods in use, while Section 4 deals with the methodology and techniques to obtain the models and also includes a description of the experimentation and a discussion of the results. Finally, the conclusions are presented and future lines of work outlined.

2. A real case study: a plastic products factory

This investigation examines the MES of a plastic products factory in Spain that manufactures different products: tubes, (polypropylene) sheets and (garbage) bags, among others. Its production process is divided into a storage area, an extrusion area and a printing and clothing area.

Figure 1 depicts the schema of the plastic bags factory where the production system is totally supervised and monitored.

Each machine has its own control system based on Programmable Logic Controllers (PLC). There are up to 75 machines, each producing a range of different products. There are also several Human Machine Interfaces (HMI) connected to an Ethernet network and a Data Acquisition System (DAQ) that collects various process signals, pressures and temperatures, among others.

The operators control the machines that are programmed to manufacture the product. Finally, the monitoring and supervisory computers that are connected to the network request information from the PLCs and DAQs. The entire network is known as the Manufacturing Control System (MCS).

A real data set was compiled with the aim of discovering new knowledge from data that could be integrated in the MES system. The main idea is to improve the capacity of the staff to estimate operational costs and production scheduling in the factory. Only a small amount of data was available, as the company had only recently begun to store data in a database management system.

Firstly, the production dynamics characteristics should be determined to integrate the MES into this scenario. To do so, the available data set has to be pre-processed and its relevant variables and partitions should be extracted according to manufacturing conditions. Once the manufacturing dynamics data have been pre-processed, a production operations model may be formulated [25].

2.1. Analysis of the objectives

As stated before, the final objective of this study is to obtain hidden know-how in the data set and to show how it can be incorporated as one or more models in the real MES to support factory staff in tasks relating to budgeting – costs management- and operator scheduling. Consequently, all available data sets in the MES from the MCS should be examined in the design of the final database.

The working method had been analysed in order to propose a solution for the first objective. Although there is a product catalogue, the client can also order customized products. When a client orders a product, its characteristics are all defined, i.e.: gross material usage, product specifications etc. Then a staff member analyses the requirements, assigns the job to a certain machine chain and estimates its cost. This process is as yet not automated, so the employee needs to analyze several plots and reports before assigning a machine chain. The challenge throughout these steps is to develop a classifier for the cost level that is associated with the configuration of a product that will be manufactured, the client’s specifications and the proposed machine. This will allow bidding in accordance with the real cost of the productive operations.
A further problem to address is the assignment of operators for machining certain products, so that staff can decide on the best shift schedules.

In this case, the inputs to the model should be the product, client, machine and operator identifiers, among others. The challenge is to classify the effectiveness of the configuration.

The greater the interpretability of the models, the better prepared it is to perform to a high degree of accuracy. This may be demonstrated through a comparison with non-interpretable models.

One of the difficulties in developing a broad solution to this problem is that the literature only contains different ad hoc solutions to specific problems [18, 25, 27, 28, 90]. A solution may be proposed based on these ideas, but it is not a clear extension of previously published works which would satisfy the objectives of the study.

The complex task of integrating the different systems should consider open architectures and clearly defined procedures for interchanging information. These procedures are currently available as open standards; different structures for the design and integration of the MES are discussed in [27].

The problem of integrating the MES, the data warehouse, online analytical processing and data mining systems have previously been discussed in [25], where decision trees were used to extract, learn and model the required knowledge. In [18], integrations are proposed in which the customer’s system must not have access to the MES data directly. Consequently, once suitable models were chosen, their integration within the MES should consider the above-mentioned ideas. The analysis of the data and the study of suitable models is the main purpose of this research, while the selection of suitable models and their integration within the MES is left for a future occasion.

3. Prototyping and Experimentation

Having collected the available data set, several tasks should be performed. Firstly, the data set has to be analyzed and pre-processed in order to determine whether there are any dependent variables. An analysis is also necessary to decide whether the data should be normalized, and whether a partition of the class variable is needed.

As stated in previous sections, two problems had to be solved and a model for each task should be obtained: a model for assisting with budgeting and a model for shifting the operators to each machine.

This Section refers to all the tasks that are needed to obtain the models for both problems. Firstly, data harvesting is outlined. Subsection 3.2 refers to data mining and the knowledge extraction tools and techniques that are proposed, while Subsection 3.3 looks at data set pre-processing. Subsections 3.4 and 3.5 describe the modelling of the budgeting problem and the shift work rota and scheduling problem, respectively. Finally, the results are discussed in Subsection 3.6.

3.1. Data acquisition from the MCS

The MCS framework is the source of the data to develop the modelling tasks. In this study, data availability is restricted to data that is currently available to staff with a decision-making role. This restriction is because the effectiveness of these models should be evaluated with the relevant staff, who might otherwise draw their own conclusions from the results.

Consequently, the available data is fixed to a predetermined feature set, which includes the following variables:

- Client: the name of the company that requests the manufacture of certain products, which generates a manufacturing order number for the product.
- Product: identification of the product that will be manufactured.
- Machine: the identifier of the machine that was assigned to a manufacturing order.
- Day: the date when manufacturing began.
- Operator: the operator or operator team that manufactures a product per day and per machine. This is a non-atomic field.
- Units: the amount of product units manufactured by a machine for a manufacturing order.
- Kg. of units produced: total units manufactured in kg.
- Discarded units: the number of discarded product units in a machine. Units discarded represent quality errors.
- Kg. of discarded units: manufactured units discarded in kg.
- OEE: overall equipment effectiveness index.
- Time spent in production: the sum of the operating time (Tm) –the run time-, the stoppage time (Ts) and micro-stoppage time (Tµs) of the machine. -
- Run time (Tm): the total time spent manufacturing a product.
• Stoppage time (Ts): total machine stoppage time during product manufacturing.
• Micro-stoppage time (Tµs): total micro-stoppage time during the manufacture of the product.
• Stoppages: the number of stoppages that take place during the manufacture of the product.
• Micro-stoppage: the number of micro-stops during the manufacture of the product.

3.2. Data mining and Knowledge extraction issues

Knowledge Extraction based on Evolutionary Learning (KEEL) software [17] was used in all the experimental and modelling stages. KEEL software is a research and educational tool for modelling data mining problems which implements more than one hundred algorithms, including classification, regression, clustering, etc.

Moreover, it includes data pre-processing and post-processing algorithms, statistical tests and reporting facilities. Finally, it has a module for data set analysis and formatting, which was used for the first task in this experiment.

There are several data mining and knowledge extraction tools, such as the Weka [45] and the Orange [51] suites, all of which could be valid for this experiment. An exhaustive list of these kinds of suites can be found in [66].

Several different techniques were selected to extract knowledge from the data set gathered from the MES. Fuzzy Rule-Based Systems and Decision Trees are considered suitable for IT support tools because of the interpretability of their models [69, 120]. Several techniques are also able to manage the type of data that is available.

Different techniques were used to compare the results and the viability of the models. The statistical methods included Quadratic Discriminant Analysis (QDA) [75], the Multinomial Logistic regression model with a ridge estimator (LOG) [24], the Kernel Classifier [75], and the K-nearest neighbour [33].

The fuzzy rule-based methods included the Fuzzy Adaboost rule learning method (ADA) [53], the Fuzzy GA-P algorithm (FGAP) [93] and the Ishibuchi Hybrid Fuzzy GBML (HFG) [52]. Finally, two well-known decision tree and decision tree rule-based methods were used: the C4.5 [84] and C4.5 rule-based methods. (C45R) [85].

In the QDA algorithm, the cost of classifying an example X with class k is calculated through Eq. (1), where \( \pi \) is the unconditional prior class \( k \) probability estimated from the weighted sample, and \( \mu \) and \( \Sigma \) are, respectively, the mean population vector and the covariance matrix for the \( k \) class. Hence, an example \( X \) is assigned with the minimum cost class as stated in Eq. (2).

\[
d_k(X) = (X - \mu_k)^T \Sigma_k^{-1} (X - \mu_k) + \ln|\Sigma_k| - \frac{1}{2} \cdot \ln \pi_k
\]

\[
d_k = \min_{l \in \{1,...,K\}} d_l(X)
\]

The LOG algorithm is based on the standard logistic regression. The probability that class \( k \) correctly classifies the example \( X = \{X_1, ..., X_p\} \) is calculated following Eq. (3), where the parameter \( \beta = \{ \beta_1, ..., \beta_p \} \) is estimated, i.e., with the maximum likelihood estimation obtained by maximizing Eq. (4). The class with higher probability is chosen to label the example.

\[
p(k|X) = \frac{\exp(\sum_{j=1}^{p} \beta_j X_j)}{1 + \exp(\sum_{j=1}^{p} \beta_j X_j)}
\]

\[
\ell(\beta) = \sum_{k} \left( k \cdot \log p(k|X) + \log(1 - p(k|X)) \right)
\]

The Kernel method is a Bayes rule classifier that, as stated in [38], uses a “non-parametric estimation of the density functions through a Gaussian kernel function.” In the KEEL software, an ad-hoc method performs covariance matrix tuning. In contrast, the K-nearest neighbour method classifies the example \( X \) with the majority class in \( K \) examples of the data set at the shortest distance from \( X \). Note that the use of the KNN implies that a metric is defined in the space to measure the distance between examples.

The Fuzzy Adaboost method is based on boosting N weak fuzzy classifiers (that is, N unreliable fuzzy classifiers are weighted according to their reliability) so that the whole outperforms each of the individual classifiers. Moreover, each example in the training data set is also weighted and tuned in relation to the evolution of the whole classifier.

The GAP is a Fuzzy Rule-Based Classifier trained using the Genetic Programming principles but using the Simulated Annealing algorithm to mutate and to evolve both the structure of the classifier and the parameters. The whole Fuzzy Rule set will evolve in each iteration.
The Ishibuchi Hybrid Fuzzy Genetic Based Machine Learning method represents a Pittsburgh style genetic learning process which is hybridized with the Michigan style evolution schema: after generating the \((N_{pop}-1)\) new Fuzzy Rule sets, a Michigan style evolutionary scheme is applied to each of the rules for all the individuals. Recall that each individual is a complete Fuzzy Rule set.

Finally, the C4.5 algorithm is a well-known decision-tree method based on information entropy and information gain. A node in the decision tree is supposed to discriminate between examples of a certain class based on a feature value. At each node, the feature that produces the higher normalized information gain is then chosen. In the case of C4.5R, the decision tree is presented as rules, where each node in the path from the root to a leaf is considered an antecedent of the rule. These rules are then filtered to eliminate redundant or equivalent ones.

Fig. 1. A schematic diagram of the MES installed in the plastic products factory. The PLCs which control each machine and the DAQs and HMIs connected through the field network constitute the MCS.

3.3. Pre-processing and partitioning of the data set

The budgeting problem and the shift work rota and scheduling problem require different data-set pre-processing and partitioning steps.

For the budgeting problem, a data set was gathered from the IT framework that amounted to 1,471 examples, containing the available historical records that comprised 22 input variables and included the features mentioned in Subsection 3.1.

After analyzing the original data set it was found that most of the examples corresponded to the tuning of the plant, and could therefore be discarded. In addition, a large quantity of totally erroneous samples were also found, which also had to be discarded.

The output variable is a class variable that represents the level of the production cost. The staff at the company chose the two different partitioning schemes. The first partition is a three-class problem, with the labels \{Low, Medium, High\}.

In the second partition scheme, the problem is divided in a two-class problem, with the labels \{Low, \neg Low\}. The examples classified as \neg Low have also been classified as \{Medium, High\}. In this second partitioning scheme it is assumed that two classifiers should be obtained: the first one discriminates the
Low class and a second (only obtainable with the \(\neg\text{Low}\) examples) discriminates the Medium class.

Several relationships were found between the features of the data set; i.e.: the one between the number of faulty units and the weight of discarded material.

Consequently, the original data set was reduced from 1,471 examples to the final data set containing 168 examples. This final data set has 4 input variables: product identification, machine identification, client identification and the number of units to produce.

In the case of the shifting and scheduling problem the original data set gathered from the IT framework contained 2,792 examples, including historical records of up to 30 input variables. The same two partitioning schemes used for the budgeting problem were applied to the scheduling problem. Once again, the staff proposed the labels for each example. The labels for the scheduling problem are \{Good, Compromise, Invalid\} and \{Good, \(\neg\text{Good}\)/\{Compromise, Invalid\}.

The final data set, after pre-processing and filtering and the relationships discovery process, contained 1215 samples with 2 input variables.

3.4. Modelling the budgeting problem

The objective of this model is to determine whether an association between a product and a specific machine will generate a certain cost level.

The methods described in the subsection 3.1 were used to obtain the classifiers. The two partitioning schemes were analyzed and modelled for each case. As there were so few examples, a 10-fold cross-validation scheme was selected and performed in a KEEL environment. The classifying error was used as the index for choosing the best algorithm.

The parameters of each method are shown in Table 1, alongside their corresponding acronyms.

<table>
<thead>
<tr>
<th>Method</th>
<th>Computer-based model for budgeting</th>
</tr>
</thead>
<tbody>
<tr>
<td>Kernel Classifier (KC)</td>
<td>Sigma Kernel = 0.01 (KC01) or 0.05 (KC05)</td>
</tr>
<tr>
<td>K nearest Neighbour (KNN)</td>
<td>Distance function: Euclidean; K value = 1 (KNN1) or 3 (KNN3)</td>
</tr>
<tr>
<td>Multinomial Logistic regression</td>
<td>Ridge value= (10^{-8}); iterations=1</td>
</tr>
<tr>
<td>model with a ridge estimator (LOG)</td>
<td></td>
</tr>
<tr>
<td>Fuzzy AdaBoost (ADA)</td>
<td>Number of labels = 3; number of rules = 8;</td>
</tr>
<tr>
<td>C4.5</td>
<td>Pruned; confidence = 0.25; 2 instances per leaf</td>
</tr>
<tr>
<td>Rule based C4.5</td>
<td>Threshold = 10; confidence = 0.25, 2 instances per leaf</td>
</tr>
<tr>
<td>Ishibushi Hybrid Fuzzy GBML (HFG)</td>
<td>number of fuzzy rules = 35; number of fuzzy rule sets = 200; crossover probability = 0.9; 1,000 generations; probability of Michigan iteration = 0.5</td>
</tr>
<tr>
<td>Fuzzy GAP (GAP)</td>
<td>number of labels = 7; number of rules = 35; population size = 50; 2 islands; iterations = 1,000; tournament size = 4; mutation and migration probabilities = 0.01 and 0.001; mutation amplitude = 0.1; 8 niches; intra-niche migration, gp and ga crossover and mutation probabilities = 0.75, 0.5 and 0.5; length of the gap chain = 10; tree height = 8</td>
</tr>
</tbody>
</table>
As may be seen, the best results are obtained for the Kernel and the Fuzzy AdaBoost methods.

The results from the two-class partitioning scheme experiments are presented in Table 2, Figure 2 and Figure 3. As may be seen, the best models were the Kernel method and Fuzzy AdaBoost, which are capable of correctly classifying more than 90% of the actual samples. The drop in algorithm performance may be due to the small size of the data set. Moreover, it appears that greater effort may be needed at the pre-processing stage, i.e.: discretization of all variables, analysis of different missing values and techniques, etc.
Fig. 3. The budgeting computer-based model: Experimental results for the {Medium, High} two-class problem.

The best results were obtained by the Kernel and the Fuzzy AdaBoost methods.

Table 3
The computer-based model for budgeting: mean results obtained for the classifiers in the {Low, Medium, High} three-class experiments. GCE: Global Classification Error; SGCE: standard deviation of the GCE; CC: percentage of correctly classified examples.

<table>
<thead>
<tr>
<th>Method</th>
<th>GCE</th>
<th>SGCE</th>
<th>CC</th>
</tr>
</thead>
<tbody>
<tr>
<td>C4.5</td>
<td>0.2974</td>
<td>0.0441</td>
<td>0.7026</td>
</tr>
<tr>
<td>C4.5R</td>
<td>0.3103</td>
<td>0.0967</td>
<td>0.6897</td>
</tr>
<tr>
<td>KC01</td>
<td>0.1077</td>
<td>0.0648</td>
<td>0.8922</td>
</tr>
<tr>
<td>KC05</td>
<td>0.1077</td>
<td>0.0531</td>
<td>0.8923</td>
</tr>
<tr>
<td>KNN1</td>
<td>0.3445</td>
<td>0.0796</td>
<td>0.6555</td>
</tr>
<tr>
<td>KNN3</td>
<td>0.3684</td>
<td>0.1120</td>
<td>0.6316</td>
</tr>
<tr>
<td>LOG</td>
<td>0.2434</td>
<td>0.0840</td>
<td>0.7566</td>
</tr>
<tr>
<td>QDA</td>
<td>0.3338</td>
<td>0.0857</td>
<td>0.6662</td>
</tr>
<tr>
<td>FGAP</td>
<td>0.4118</td>
<td>0.0975</td>
<td>0.5975</td>
</tr>
<tr>
<td>ADA</td>
<td>0.1783</td>
<td>0.0785</td>
<td>0.8217</td>
</tr>
<tr>
<td>HFG</td>
<td>0.3857</td>
<td>0.0799</td>
<td>0.6143</td>
</tr>
</tbody>
</table>

Fig. 4. The budgeting computer-based model: Experiment results for the {Low, Medium, High} three-class problem.

The best results were obtained by the Kernel methods, then by the Fuzzy AdaBoost method.

3.5. Modelling the scheduling problem

The aim of this model is to determine whether an association between a product and a specific machine and operator would be suitable.

The methods described in Subsection 3.2 have been used again to obtain the classifiers; Table 1 shows the parameters used for each of the methods, along with their corresponding acronyms.

The two partitioning schemes were also analyzed and modelled for each case. The 10-fold cross-
validation schema was selected and performed in a KEEL environment. Again, the classifying error was used as the index to choose the best algorithm.

**Table 4**
The scheduling problem model: mean results for the classifiers in the \{Good, ¬Good\} two-class experiments.

<table>
<thead>
<tr>
<th>Method</th>
<th>GCE</th>
<th>SGCE</th>
<th>CC</th>
<th>GCE</th>
<th>SGCE</th>
<th>CC</th>
</tr>
</thead>
<tbody>
<tr>
<td>C4.5</td>
<td>0.3166</td>
<td>0.0032</td>
<td>0.6834</td>
<td>0.4453</td>
<td>0.0016</td>
<td>0.5547</td>
</tr>
<tr>
<td>C4.5R</td>
<td>0.6555</td>
<td>0.3445</td>
<td>0.0340</td>
<td>0.4935</td>
<td>0.0553</td>
<td>0.5065</td>
</tr>
<tr>
<td>KC01</td>
<td>0.3166</td>
<td>0.0032</td>
<td>0.6834</td>
<td>0.4453</td>
<td>0.0016</td>
<td>0.5547</td>
</tr>
<tr>
<td>KC05</td>
<td>0.3166</td>
<td>0.0032</td>
<td>0.6834</td>
<td>0.4453</td>
<td>0.0016</td>
<td>0.5547</td>
</tr>
<tr>
<td>KNN1</td>
<td>0.6456</td>
<td>0.0347</td>
<td>0.3544</td>
<td>0.5472</td>
<td>0.0115</td>
<td>0.4573</td>
</tr>
<tr>
<td>KNN3</td>
<td>0.5757</td>
<td>0.0365</td>
<td>0.4243</td>
<td>0.5295</td>
<td>0.0383</td>
<td>0.4705</td>
</tr>
<tr>
<td>LOG</td>
<td>0.3166</td>
<td>0.0032</td>
<td>0.6834</td>
<td>0.4537</td>
<td>0.0182</td>
<td>0.5463</td>
</tr>
<tr>
<td>QDA</td>
<td>0.3166</td>
<td>0.0032</td>
<td>0.6834</td>
<td>0.4669</td>
<td>0.0313</td>
<td>0.5331</td>
</tr>
<tr>
<td>FGAP</td>
<td>0.3224</td>
<td>0.0253</td>
<td>0.6776</td>
<td>0.4489</td>
<td>0.0348</td>
<td>0.5511</td>
</tr>
<tr>
<td>ADA</td>
<td>0.3166</td>
<td>0.0032</td>
<td>0.6834</td>
<td>0.4573</td>
<td>0.0148</td>
<td>0.5427</td>
</tr>
<tr>
<td>HFG</td>
<td>0.3232</td>
<td>0.0165</td>
<td>0.6768</td>
<td>0.4670</td>
<td>0.0445</td>
<td>0.5330</td>
</tr>
</tbody>
</table>

**Table 5**
The scheduling problem model: mean results for the classifiers in the \{Good, ¬Good\} and the \{Compromised, Invalid\} two-class experiments.

<table>
<thead>
<tr>
<th>Method</th>
<th>GCE</th>
<th>SGCE</th>
<th>CC</th>
<th>GCE</th>
<th>SGCE</th>
<th>CC</th>
</tr>
</thead>
<tbody>
<tr>
<td>C4.5</td>
<td>0.5994</td>
<td>0.0376</td>
<td>0.4006</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
The results from the first experiment are presented in Table 4, Figure 5 and Figure 6. As may be seen, it correctly classifies almost the 70% of the examples, when discriminating between the Good and ~Good classes. In the second three-class problem, the results obtained are quite discouraging (see Table 5) as explained in the next sub-section and, consequently, no graphics results are shown.

3.6. Discussion

The results from the scheduling problem are not as suitable as expected, mainly due to the low quality of the data; in other words, the information contained in the data is not informative enough. One of the reasons for this behaviour emerged when presenting the results to the company: the examples in the data set were not independent samples.

The case of shift work rotas with two operators to manufacture a product on the same machine was treated as two examples with the same data, but they involve different operator identification information. Consequently, if one of the two operators performed better than the other, there was no way to distinguish between them. Unfortunately, this was quite common in the data set, which implied poorer results for the classifiers.

An important conclusion may be drawn from this experiment: the data set should be more informative and representative of the problem, if better models have to be generated. The company should rely on an in-depth analysis of the available data and measurements, but it is also necessary to study the relationships between the variables under consideration, i.e. using Cooperative Maximum Likelihood Hebbian Learning (CMLHL) [30] as shown in [102, 103].

For the case of the budgeting problem, the above mentioned Fuzzy Ada-Boost and the Kernel methods are the more suitable algorithms. Improvements to the FGAP and the HFG algorithms are also suggested, in order to reduce the standard deviation.

It could be said that these two methods may improve their performance with a better definition of their parameters (population and sub-population sizes, number of islands, etc.) and if they had a larger number of generations.

It is worth remarking that this is especially interesting in the case of a two-class problem, the performance of which is better than it is in the three-class problem.

Obviously, there is not enough information to obtain good models in the budgeting problem for the three-class problem.

As an example, the number of units to be produced is somewhat dependent on the machine, as each machine has a maximum production rate. But this data was not used in the experiment, so it was not possible to normalize those variables, which in turn, reduced the quality of the classifier.

If a model should be chosen for the shifting and scheduling problem, it could be said that the interpretable models represent the best option as they all perform with a similar error distribution but they allow the staff to analyse their proposed decisions. In the case of the shifting and scheduling problem, results in Table 4 show that the C4.5 and the kernel methods are the most suitable for the first problem (the computer-based support model for budgeting), as presented in [25]. Nevertheless, the relatively low accuracy of all the algorithms suggest that we redesign the way the data set is formed and repeat all of the modelling steps. The improved data set, with larger amounts of information on the process, will surely enhance the algorithms results.

4. Conclusions

This interdisciplinary research presents how to analyze and discover knowledge within the data set in an MES for its integration in the system, to improve factory capacity and support the staff of a plastic products factory with budgeting costs management and shift work rotas for operators.

A classification model for the budgeting problem was solved through the application of different com-

<table>
<thead>
<tr>
<th>Classifier</th>
<th>Accuracy</th>
<th>Precision</th>
<th>Recall</th>
</tr>
</thead>
<tbody>
<tr>
<td>C4.5R</td>
<td>0.6357</td>
<td>0.0324</td>
<td>0.3643</td>
</tr>
<tr>
<td>KC01</td>
<td>0.6209</td>
<td>0.0025</td>
<td>0.3791</td>
</tr>
<tr>
<td>KC05</td>
<td>0.6160</td>
<td>0.0160</td>
<td>0.3840</td>
</tr>
<tr>
<td>KNN1</td>
<td>0.6883</td>
<td>0.0147</td>
<td>0.3117</td>
</tr>
<tr>
<td>KNN3</td>
<td>0.6917</td>
<td>0.0205</td>
<td>0.3083</td>
</tr>
<tr>
<td>LOG</td>
<td>0.6308</td>
<td>0.0247</td>
<td>0.3692</td>
</tr>
<tr>
<td>QDA</td>
<td>0.6250</td>
<td>0.0216</td>
<td>0.3750</td>
</tr>
<tr>
<td>FGAP</td>
<td>0.9416</td>
<td>0.0241</td>
<td>0.0584</td>
</tr>
<tr>
<td>ADA</td>
<td>0.6226</td>
<td>0.0278</td>
<td>0.3774</td>
</tr>
<tr>
<td>HFG</td>
<td>0.6118</td>
<td>0.0324</td>
<td>0.3882</td>
</tr>
</tbody>
</table>
puting techniques. The selected models are able to classify the different levels of costs. However, the scheduling problem cannot be solved with the same approach and the same initial data collection.

It is worth remarking on the importance of analyzing the data that has to be gathered before performing the experiments. The example of the scheduling problem clearly illustrates this sort of situation.

An experimental design about a clear list of the objectives to be accomplished by the MES should be prepared prior to the collection and analysis of relevant data.

Future work will include modelling using different heuristics, i.e. multi-objective solutions and simulated annealing, with the aim of solving problems that can be integrated within the enterprise resource planning. More data should be gathered from the plant, such as machine operating limits and a full experiment is need to establish how a group of operators can cause fluctuations in the level of the OEE index. Finally, prior analysis of the data through the use of well-known techniques (such as CMLH) would contribute to the evolution of MES design and engineering.

Acknowledgements

This research has been partially funded by the Spanish Ministry of Science and Innovation, under project TIN2008-06681-C06-04 and TIN2010-21272-C02-01, the Spanish Ministry of Science and Innovation through project PID 560300-2009-11, and project BU006A08 and CCTT/10/BU/0002 of the Junta de Castilla y León. The authors would also like to thank the manufacturer of vehicle interior components, Grupo Antolin Ingeniería, S.A., as part of the MAGNO 2008 - 1028.- CENIT Project funded by the Spanish Ministry of Science and Innovation.

References

[22] L. Carro-Calvo, S. Salcedo-Sanz, R. Gil-Pita, A. Portilla-Figuera and M. Rosa-Zurera. An Evolutive Multiclass Algorithm for Automatic Classification of High Range Resolu-
[51] Laboratory of Artificial Intelligence, University of Ljubljana, Orange Machine Learning Library, http://orange.biolab.si/.


