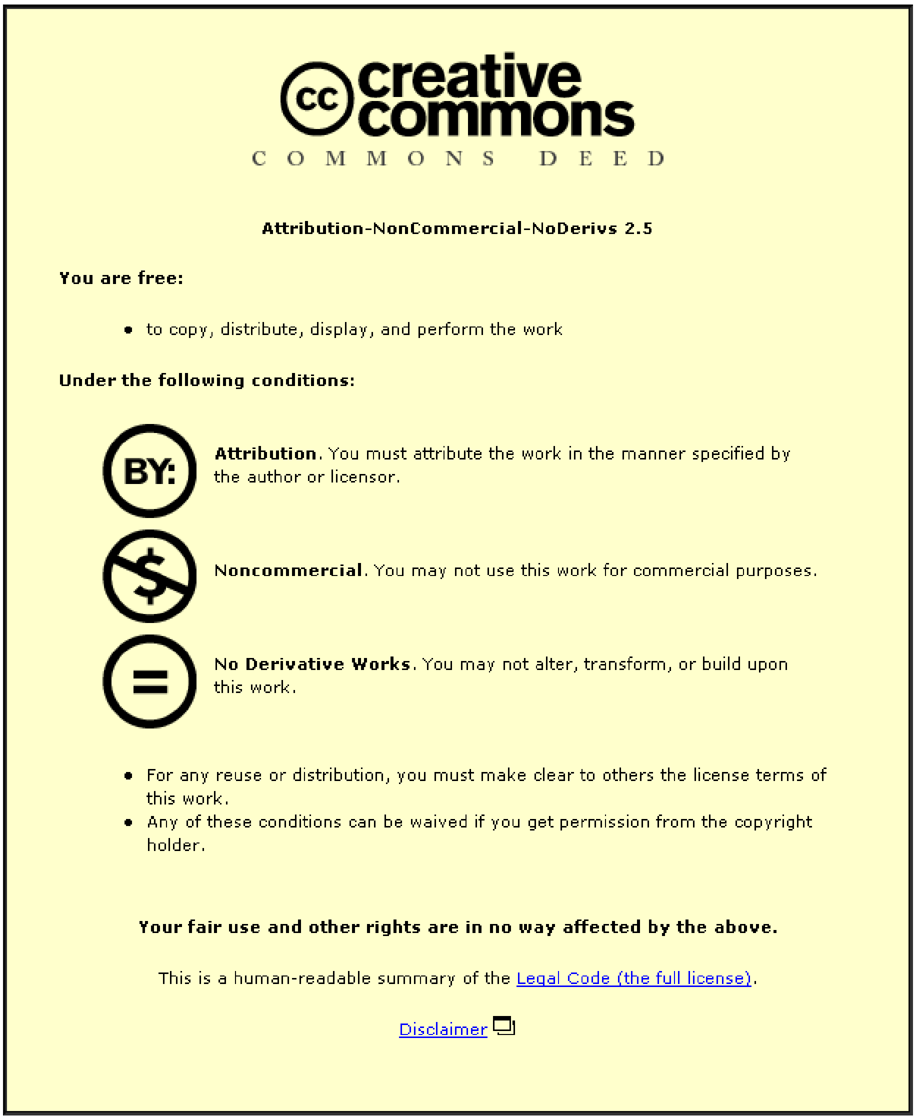

Product design and manufacturing process improvement using association rules

M Shahbaz[1], **Srinivas**[1], **J A Harding**[1]*, and **M Turner**[2]

[1]Wolfson School of Mechanical and Manufacturing Engineering, Loughborough University, Loughborough, UK
[2]Rolls Royce plc, Barnoldswick, UK

The manuscript was received on 15 April 2005 and was accepted after revision for publication on 3 October 2005.

DOI: 10.1243/095440506X78183

Abstract: Modern manufacturing systems equipped with computerized data logging systems collect large volumes of data in real time. The data may contain valuable information for operation and control strategies as well as providing knowledge of normal and abnormal operational patterns. Knowledge discovery in databases can be applied to these data to unearth hidden, unknown, representable, and ultimately useful knowledge. Data mining offers tools for discovery of patterns, associations, changes, anomalies, rules, and statistically significant structures and events in data. Extraction of previously unknown, meaningful information from manufacturing databases provides knowledge that may benefit many application areas within the enterprise, for example improving design or fine tuning production processes. This paper examines the application of association rules to manufacturing databases to extract useful information about a manufacturing system's capabilities and its constraints. The quality of each identified rule is tested and, from numerous rules, only those that are statistically very strong and contain substantial design information are selected. The final set of extracted rules contains very interesting information relating to the geometry of the product and also indicates where limitations exist for improvement of the manufacturing processes involved in the production of complex geometric shapes.

Keywords: knowledge discovery in databases, data mining, manufacturing, association rules, product data, product quality and quality control

1 INTRODUCTION

Market pressures for improved quality and faster responses within the supply chain have resulted in increased measurement and recording of product and process data throughout the product life cycle. Information is also routinely recorded to satisfy component traceability requirements and quality audits. Therefore, over time, the manufacturing industry captures and stores huge amounts of detailed data. These data may be related to designs, product manufacture, workstation operation and scheduling, processes and performance, use of particular machinery and other resources, sales, inventory, and marketing. Hence a vast range of data is routinely collected as part of the everyday operation of any manufacturing enterprise. However, these collected data are commonly not exploited to their full potential. Operational data may be used for business reports and simple statistical analyses; to provide information about where the enterprise is currently standing against its past performance or to enable comparison to be made between the enterprise and its competitors. Hence, this potentially valuable knowledge resource is generally not thoroughly understood, reused, or exploited. In recent years, efforts have been made to utilize the existing databases from manufacturing enterprises for design and quality control processes using, for example, the factory data model [**1**, **2**] and data warehouses. However, manufacturing's largely unexplored data sources need to be thoroughly understood before their embedded

**Corresponding author: Wolfson School of Mechanical and Manufacturing Engineering, Loughborough University, Ashby Road, Loughborough LE11 3TU, UK. email: j.a.harding@lboro.ac.uk*

knowledge can be properly exploited as explicit knowledge.

The current limited use of the accumulated data has led to the 'rich data but poor information' problem. Existing databases generally contain large numbers of records and attributes that need to be simultaneously explored because of the possible relationships and interrelationships that may exist. The volume and/or complexity of such data generally make manual analysis impossible and therefore automated analysis is essential for knowledge to be extracted in a form that can benefit the business. Hence, knowledge discovery in databases (KDDs) and data mining methodologies will become extremely important tools in future manufacturing systems. Data mining is an emerging area of computational intelligence that offers new theories, techniques, and tools for the analysis of large data sets. Data mining is the search for valuable information in large volumes of data, and has been defined as 'the non-trivial process of identifying valid, novel, potentially useful and ultimately understandable patterns in data' [**3**].

The retrieved knowledge has even greater value if it can be integrated with organizational strategies to improve the decision-making process [**4**, **5**]. However, it is very difficult to design a data-supported manufacturing enterprise that can benefit from its historical or legacy systems as well as from its current databases. To date, little has been done to exploit the potentially viable data mining technology in manufacturing industries, and indeed literature on the application of data mining in manufacturing has only recently appeared, for example references [**6**] to [**9**].

1.1 Motivations

Since large stores of data already exist in manufacturing enterprises, it is not clear why technologies such as KDDs are not commonly used in the engineering industry. It may be because of the long time scales and expense involved in introducing these new techniques, or it could be because of the uncertainty of payback values. An alternative possible reason might be the complexity and diversity of different manufacturing processes, as these can make it extremely difficult to devise a generic data mining process that can be used for all kinds of manufacturing processes and can tackle all types of manufacturing problems. However, implementation of data mining technology in other areas such as banking, finance, marketing, insurance, telecommunication, health care, etc. has given very good results [**10**, **11**] and people are already benefiting from the knowledge discovered in these fields.

The research reported in this paper is based on the belief that the explicit knowledge and information extracted from existing data warehouses, or from current product and production processes, can be used to improve the performance of manufacturing systems. Indeed, recent dramatic decreases in acceptable timescales for product development cycles and increases in the automation of manufacturing processes are forcing planners to identify and make use of knowledge that may be hidden in their existing and historical data sources. This paper will show that explicit knowledge that has been extracted using data mining techniques, can be used to improve the design or redesign of products and to identify manufacturing process constraints.

The manufacturing industry relies on complex systems and machines. Operational efficiency, reliability, and cost have all been improved by the skill of designers and the use of a variety of analytical, computational, and manufacturing techniques. Managers and designers may receive feedback from manufacturing operators as to what is effective and what is not so effective, but this tends to be based on anecdotal experience. The overall objective of the work presented here is to use emerging data mining techniques to provide this feedback in a more formalized manner. This work is part of an ongoing research programme that has the overall objective of providing a methodology and information system design to support explicit knowledge acquisition and reuse through the application of data mining technology in manufacturing industries. The results reported here are from the first phase of the research, which has focused on the area of improving product and process understanding through data mining.

This research has been actively supported by Rolls Royce plc, and uses case studies based on the complex manufacturing processes involved in fan blade manufacture (see Fig. 1). For reasons of confidentiality and to set this research into context, a simplified manufacturing scenario will be considered, and this is now briefly described using Fig. 2, which shows the manufacturing system as a block diagram, with a sequence of processes identified as process 1 to process *n*. The parameters of the different manufacturing processes may be measured at each individual station or machine and the dimensions or quality of the product may also be measured with coordinate measuring machines (CMMs) after every important step.

The example in Fig. 2 shows data being extracted from two processes. The earlier process generates data related to different parameters of the machines and these might include, for example revolutions per minute, temperatures, cycle time, pressure, tool changes, down times, etc. The second data extraction point may be a metrology process, which measures different dimensions of the product, for quality control purposes.

Each manufacturing stage is important and contributes to the final results and required quality of the product, but some manufacturing processes and their parameters may be more influential than others. In the real, complex manufacturing system, it is very difficult to determine the overall effect of particular parameters of a certain manufacturing process on the final quality of the product. However, this is essential if the process is to be effectively controlled and the desired product quality and throughput achieved. Further complexities may exist due to any unknown (or unconfirmed) interrelationships between dimensions and parameters of the product.

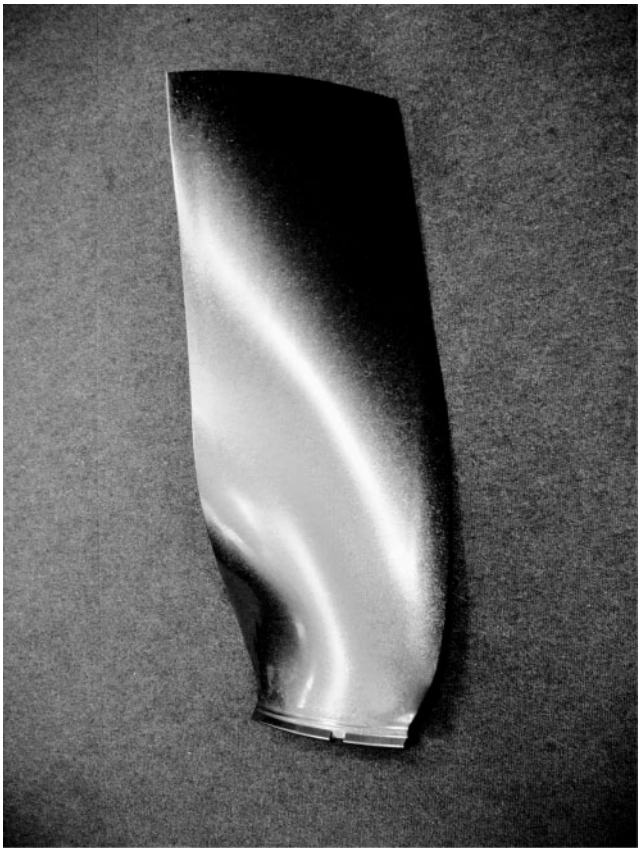

Fig. 1 A Trent 500 wide-chord fan blade, manufactured by Rolls Royce plc

This paper examines association rule mining and shows how it can be implemented on manufacturing data as part of a knowledge acquisition process. This is done to determine how explicit knowledge can be extracted from existing databases, so that it can be used: (1) to improve the design of the product and (2) to discover any constraints that might exist in the manufacturing system. The results obtained are valuable and have been confirmed by experts in this field. A methodology is also provided for the implementation of association rule methods in manufacturing enterprises. The reported work focuses on product dimensions. These are important as there are many complex interrelationships within the intricate geometry of a fan blade. Further work is currently in progress to determine relationships between manufacturing process attributes and the output classes of the product.

The potential benefits of this research go beyond the application of association rule methods to manufacturing. The ability to identify explicit rules which relate different characteristics (and therefore implicitly metrics of quality) from production data of manufactured products, provides the potential to reuse this discovered knowledge to improve and fine-tune the production processes involved. The discovered knowledge also provides an improved understanding of the interrelationships between characteristics of the product, which can be fed back into the design of similar products or the

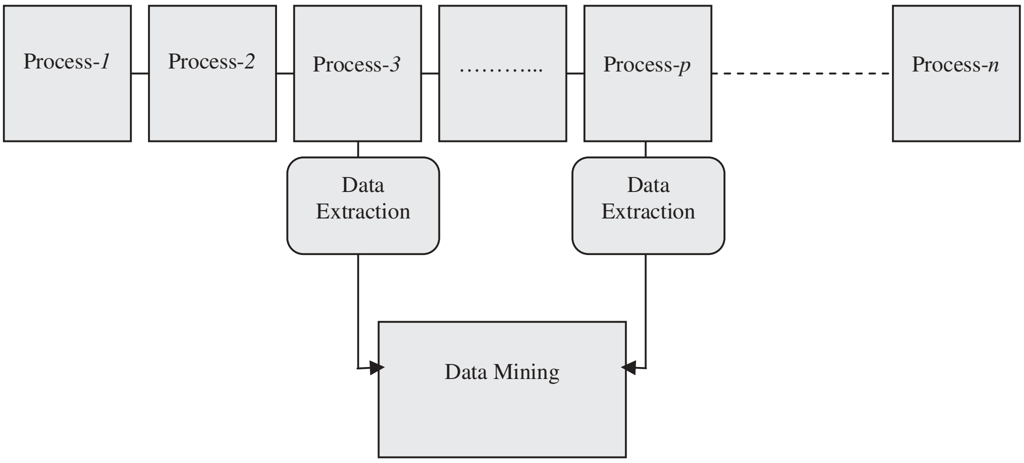

Fig. 2 A sample block diagram of a manufacturing process showing the flow of a product and data extraction

redesign or improvement of the current product. The improved understanding that is gained from the discovered knowledge is based on real production experience, and this may also be used to influence the concept and embodiment stages of new product designs. To set these benefits into context, the next section covers the relevant background of design and manufacturing.

2 BACKGROUND

Engineering design, unlike other design domains, does not directly result in a physical product. Instead it creates a set of specifications to construct or fabricate the product. Therefore, mapping design objectives and constraints to a specification is an important design issue. Most engineering design models [**12–15**] show feedback from successive design stages to earlier stages, but none of the models explicitly include feedback from life-cycle information, once a product has been turned over to manufacturing and sales [**16**]. Designers do not normally have access to product-life-cycle information and other feedback and this can cause costly design iterations and increased time to market [**16**]. Information processing-based design models, such as that proposed by Dym *et al.* in reference [**17**], do include external information files, handbooks, manuals, etc. in their mapping of design state.

Information collected during a product's life cycle provides feedback on the product's performance which can be used to assess the quality of the design [**18**]. At present the experiences and impact of this life-cycle information and trends tend to be only fed back into design through informal means, such as through some form of network, or occasionally more formally through review meetings. Figure 3 shows an aggregated and condensed version of the design and information extraction process, which indicates that information from manufacturing data may be used within the design process. Data may be recorded during manufacturing, assembly, or testing to determine, for instance, how well products generally meet certain dimensional constraints. Manufacturing process data and measurements are available from work in process (WIP) data recorded at the shop floor. These kinds of transactional data can be analysed to study the effectiveness of a design in meeting the target strength, shape, and dimensionalities. CMMs and other tests can reveal flaws in the 'form' of the product. Hence, transactional data can contain useful information that may enable designers to generate more efficient and optimal designs in the future.

In the current authors' research, data mining was used on the parameters defining the geometry of the product, to identify whether any associations were present. Association rule algorithms were therefore utilized to unearth relationships within the geometry under the dynamic behaviour of the system. The motivation for using this approach was that if explicit knowledge could be identified in this way, it should be possible to 'design in' greater control over various aspects of the geometry and to fix suitable geometric values to enhance the performance of the manufacturing.

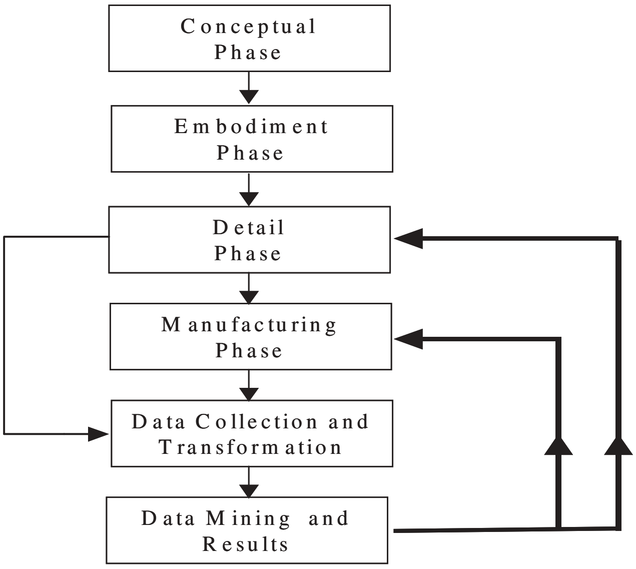

Fig. 3 Data analysis and feedback within design and manufacture

3 DATA MINING WITH MANUFACTURING DATA

Data mining is the search for hidden information, patterns, or trends in data. The whole process works in several stages including understanding the problem and process, acquisition of background knowledge, data cleaning, data selection and transformation, data mining, pattern evaluation, and knowledge representation. These stages will be examined in turn in the remainder of this section and in section 4. In this section, data cleaning and data transformation are primarily discussed in the general context of how these stages should be applied to manufacturing data. In section 4, a particular data mining process, using association rules, is described in detail. The adaptation of association rules for use on manufacturing data is a new application of association rule technology and is therefore a major contribution of this research. A worked example is provided to demonstrate the techniques that were developed during the case studies undertaken on the wide-chord fan blades. Finally, pattern evaluation is discussed by examining ways of assessing the quality of the generated rules.

3.1 Data cleaning

Data cleaning is a very important stage in the process of knowledge extraction from the database and consumes most of the resources [**19**]. Manufacturing data can be recorded in several ways as manufacturing systems may have some manual data entry systems as well as analogue and digital data acquisition systems and/or automatic data loading from computer-aided manufacturing (CAM) systems. Cleaning can be a particularly complex operation for manufacturing data, since these data are likely to contain more noise and inaccuracies than other kinds of data, such as telecommunication data, market basket data from retail purchases, or data from web logs, etc. It is therefore crucial that manufacturing data be cleaned carefully and thoroughly to make them ready for the different types of transformations that may be necessary before particular data mining techniques can be applied. Detailed examination of the whole process is particularly important when human interaction is involved at any stage of the data collection. Data that have been manually typed in by the operators may include details of the operator's ID, machine ID, starting conditions of the operation, and product input information. In such situations there is always a chance that the operator will mistype a keystroke resulting in inaccurate information being fed in. Data cleaning involves identification of such records and then correcting them if possible or otherwise removing the poor-quality records from the training and test data sets. All changes and deletions should be carefully documented, as it may be necessary to refer back to them if any anomalies are found or problems occur during the data mining process. Duplication in records is commonly a result of a machine malfunctioning and/or an operator having to restart the process on the product. Alternatively, errors also commonly occur during start-up when new products are introduced for production, as different test runs may be made on products before the actual start of the real production and this can result in multiple entries in the data.

Data integration and combination can also be a problem in manufacturing enterprises where each stage or process may record or store their individual data sets separately. Therefore, to capture the complete record of a product's journey through several processes may require product identifiers to be matched in several original sources and then consolidated in the form of a relational database. Matching the records can be a problematic task, which in practice can reduce the number of data sets considerably. There are several reasons for this, including missing data (entries not fully keyed in by the operators), lost data, data not in electronic format (for example data stored in the form of ultrasonic or X-ray images or drawings), data without the main or primary ID of the records, partial transformation of non-electronic data into electronic information, and confusing data with some kind of duplication. In such data, most of the records match, except for a few dissimilarities, which require further time and investigation. All these kinds of problems need fixing during the cleaning process. Modifications made during cleaning should also be properly recorded [**20**] for future consultation when the results need to be translated into the form of process constraints or design changes. In the current research, Microsoft Excel was used extensively for the data cleaning activities, along with ORACLE and some tailor-made C programs for particular cleaning and data transformation tasks.

3.2 Data transformation

Data transformation is not always necessary in data mining generally. However, in the case of manufacturing data it becomes a necessary step, especially when process parameters or product dimensional data are involved. Transformation is necessary because it may be virtually impossible to discover relationships for exact, continuous, measured values. The data should therefore be transformed into some appropriate, representative identifiers, before being used as input to the chosen data mining techniques, which may include decision trees, clustering, or association rules. When an acceptable model has been generated and applied, the results can be translated back into the appropriate range of variable values, using a reverse transformation process. In the case of manufacturing data the transformation stage requires detailed understanding of the manufacturing process, its constraints and operations, the importance of particular dimensions of the products, and the range of manufacturing variations that are likely. It is therefore essential for the data mining expert to work with the manufacturing engineers or process experts during this stage. Different data will need different kinds of transformation: however, if the data involve manufacturing variables then drawing the simple distribution curve and finding the standard deviation can give a good guide for transforming the data into some suitably identified ranges.

Similarly, the measured dimensions of partially manufactured or final products will also need to be transformed. The approach adopted in this research was initially to consider the current tolerance bands. The output dimensions were then divided into different sections from the upper engineering tolerance to the lower engineering tolerance as shown in Fig. 4. The reason for transforming the data into these

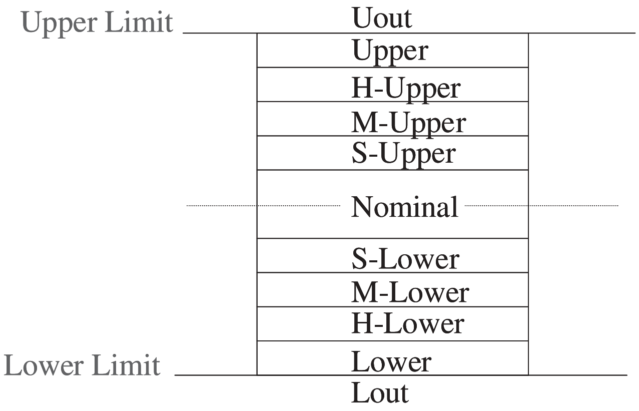

Fig. 4 Different sections/divisions for an output dimension of a product

ranges is that the manufacturing process will achieve these divisions for large batches of products, which is not the case for individual precise output dimensions. This is important as it is more appropriate to run some kinds of data mining algorithms with only a few divisions of the data since too many different values will result in very indistinct results.

4 ASSOCIATION RULES FOR MANUFACTURING: A CASE STUDY EXAMPLE

The data mining and pattern evaluation stages are discussed in this section, with particular emphasis being placed on the processes and methodology developed during the present authors' case study experiments using the dimensional data collected during the fan blade manufacture. During this research, much of the data obtained from the manufacturing process and about the products were in the form of flat files, which were cleaned to remove inconsistencies and discrepancies and then compiled into one workable table so that different data mining algorithms could be applied. The CRISP-DM (cross-industry standard process for data mining) methodology [**20**] was adopted in this case study so that the data mining techniques were applied on the manufacturing data in an appropriate, structured manner. The data used in the case studies came from complex three-dimensional aerofoil components which were produced by highly controlled manufacturing processes including diffusion bonding and superplastic forming. These processes are among the most technically sophisticated and sensitive manufacturing processes currently used in the aerospace industry. Decision trees and clustering approaches were initially tried as these data mining techniques have been the most commonly used in manufacturing studies reported in the literature [**21**]. However, these techniques did not produce sufficiently accurate and reliable results, and therefore association rule methods were then considered.

4.1 Association rule

The association rule algorithm was originally developed for very different types of data – i.e. transactional data from the retail sector, for market basket data analysis – and the core idea of association rules was presented by Agrawal *et al.* [**22**] in 1993. Agrawal and Srikant [**23**] then proposed the Apriori algorithm, for mining association rules in retail data. The association rule approach has basically remained the same since its initial presentation, however substantial research has been done to accelerate the process of mining association rules [**24–27**] in very large databases. The Apriori algorithm is now examined, as this was used in the research reported here.

The Apriori algorithm can simply be stated as, 'any subset of a large itemset must be large' [**28**]. Here, 'large' means a defined support or occurrence level of single or multiple items in the transactions.

The association or frequent itemsets are discovered by first finding the frequent occurrence of single items and then making combinations of these single items and checking whether or not the occurrence of the combination has a greater or equal defined support level. The support level is defined as the number of times an item must occur on its own or in combination with other items for it to be of interest and therefore be called a frequent itemset. The process of making combinations is repeated until the largest frequent itemsets are found. As defined earlier, the definition of the Apriori principle is that any subset of a discovered frequent itemset (call this *F*) will have the same or greater support level as *F*. Association rules can therefore be found by using all the subsets of the frequent itemsets.

4.2 Association rule mining for manufacturing data

Manufacturing is a value-adding process on raw materials since it transforms raw materials into a final product. The aim of all manufacturing processes is to produce accurate products according to the design specification. For high-precision products, high-quality manufacturing processes try to squeeze (minimize) the engineering tolerances to produce as accurately and consistently as possible.

There are many statistical methods used to control production and support the manufacture of products within design limits. However, these kinds of methods cannot identify the manufacturing limitations of the current set of production systems. An interesting result of this research is that association rule mining can be used on historical data to find any relationships that may be present, to both improve designs and give useful information about production capabilities and limitations.

Previously reported applications of association rules indicate that they have mostly been used on transactional databases or for market basket data where finding associations helps in determining the layout of displays in supermarkets or in determining what promotions should be offered. A novel aspect of this research is therefore that association rule methods have been applied to product manufacture and operational data from manufacturing industries.

A product's output dimensions are a good measure of the quality of the production cycle and can help in suggesting any alteration in the design or any dependency or relation between different dimensions resulting from the manufacturing process. It is very important for designers and production engineers to understand the interrelationships between different dimensions of the product, particularly when components have complex geometry and need to be manufactured to a high level of precision. If one dimension is positively related to another, i.e. improvements to the first also improve the second, then it may be beneficial to put resources into improving manufacturing accuracy to consistently achieve outputs within even smaller tolerance bands. In contrast, knowledge that particular dimensions are negatively related can reduce waste. It would not be practical to put additional manufacturing effort into optimizing particular features on the blade only subsequently to find that these efforts have inadvertently been detrimental to the accuracy of other aspects of the blade, perhaps tending to push another dimension out of tolerance. This type of life-cycle knowledge would also be valuable to designers on future projects. A method for identifying the associations between the output dimensions of a product, using association rules between different measures or dimensions of the products, is now demonstrated.

First, the product's dimensional data were separated from the rest of the records. The product's data had already been transformed into different categories, as shown in Fig. 4, but for convenience, they were further transformed from strings into integer identifiers to use in the Apriori algorithm. A simple nomenclature was defined to identify different output dimensions into integer identifiers. In the present authors' case studies, many dimensional values were examined for each fan blade at different sections (e.g. bow, disposition, lean, blade form angle, thickness of leading edge at different distances, wave, etc.), comprising more than 250 different sections of the blade. It is not possible to explain the proposed methodology clearly and concisely using the complex geometry of the wide-chord fan blade (see Fig. 1), and all the different variables. Therefore, to reduce the complexity and to explain the approach fully, a simplified example of a small section of the product will be demonstrated using different dimensions on a cuboid, as shown in Fig. 5.

The output dimensions of width, height, and thickness at different sections are measured by CMM and recorded. The dimension at each section has an engineering tolerance, which was divided into 11 (or any other appropriate division as convenient) sections as shown in Fig. 4. Now the measured dimensions of all the sections are translated into the appropriate bands. The data from the simplified product section are shown in Table 1, with appropriate dimensions.

The transformation table for section 'ab' of 'Thickness' is shown in Table 2. If a measured value

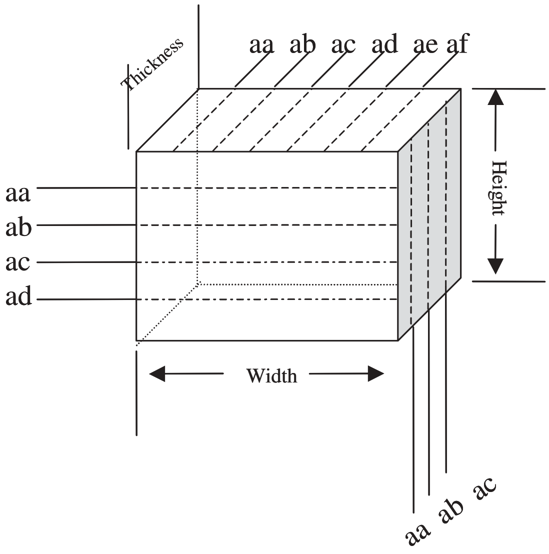

Fig. 5 A sample product with different sections

Table 1 Dimensions of different sections with tolerances

	Thickness						Width				Height		
	aa	ab	ac	ad	ae	af	aa	ab	ac	ad	aa	ab	ac
D	7.55	7.5	7.45	7.45	7.5	7.55	7.55	7.45	7.45	7.55	7.55	7.5	7.55
+	0.02	0.02	0.02	0.02	0.02	0.02	0.02	0.02	0.02	0.02	0.02	0.02	0.02
−	0.02	0.02	0.02	0.02	0.02	0.02	0.02	0.02	0.02	0.02	0.02	0.02	0.02

D, dimension.

of ab_Thickness equals 7.509 then according to the transformation it will be translated as M-Upper. This transformation is necessary since, if valid association rules are discovered, the manufacturing set-up must be able to operate between limits instead of requiring exact measured values.

4.2.1 Second transformation

The product data need to be transformed from categorized data into integer identifiers, which are easier to use as input for the association rule algorithm. This is then used to find the frequent itemsets and then the association rules.

The integer identifier transformation matrix is shown in Table 3. Each integer identifier is a combination of two, two- or three-digit numbers (these numbers can be chosen according to the requirements). The first two digits show the dimensional band and the last two digits show the section of the product. For example, if the 'ac' section of 'Width' has a measured value in the S-Lower band, then that value will be translated as 1709.

The complete set of dimensional data for each manufactured product is therefore transformed in this manner into integer identifiers, and is then treated as one transaction (i.e. forming one record, with multiple fields, in a database table). The whole training dataset (of transformed dimensional data from several hundred manufactured products) should then be fed into the association rule algorithm program. This requires the minimum acceptable support level to be selected so that the frequent itemsets can be identified in the data. Support is defined as the minimum number of occurrences of the itemsets in each stage of the iteration of the algorithm. If the support level is set too high, there is a chance that some important items in the frequent itemset data could be missed, which could lead to some valuable relationships being missed and remaining hidden. By decreasing the required level of support too far, too many 'low-quality' frequent itemsets may be found, resulting in numerous association rules, which may mostly be of little value.

Table 2 Data categorization

If	Then replace the value with
Dimension > 7.52	U-Out
7.52 > Dimension ⩾ 7.516	Upper
7.516 > Dimension ⩾ 7.512	H-Upper
7.512 > Dimension ⩾ 7.508	M-Upper
7.508 > Dimension ⩾ 7.504	S-Upper
7.504 > Dimension ⩾ 7.496	Nominal
7.496 > Dimension ⩾ 7.492	S-Lower
7.492 > Dimension ⩾ 7.488	M-Lower
7.488 > Dimension ⩾ 7.484	H-Lower
7.484 > Dimension ⩾ 7.480	Lower
Dimension < 7.480	L-Out

It is important that mining decisions are made in the context of the physical realities of the data that are being examined. In this study, the data are the result of very highly controlled manufacturing processes. Therefore the current authors' are particularly interested in unusual combinations of results that may represent occasional 'problem' situations. Hence, in this research, it was considered important not to risk missing possibly valuable hidden knowledge, simply because a particular combination of manufacturing outputs only rarely occurs. Therefore the decision was taken to keep the support level very low, and the best possible solution found was to calculate the support level by counting the occurrence of each integer identifier in the whole data and then setting the support level to equal the minimum of these counted occurrences.

Association rules were then generated from each of the frequent itemsets. Each of the frequent itemsets was split into all the possible subsets, and rules generated using these subsets, in the form: subset $x \rightarrow$ subset y. A confidence level was then calculated for each rule based on the number of times the set (subset $x \cup$ subset y) occurred in the original

Table 3 Integer identifier transformation matrix

		U-Out	Upper	H-Upper	M-Upper	S-Upper	Nominal	S-Lower	M-Lower	H-Lower	Lower	L-Out
		11	12	13	14	15	16	17	18	19	20	21
aa_Thickness	01	1101	1201	1301	1401	1501	1601	1701	1801	1901	2001	2101
ab_Thickness	02	1102	1202	1302	1402	1502	1602	1702	1802	1902	2002	2102
ac_Thickness	03	1103	1203	1303	1403	1503	1603	1703	1803	1903	2003	2103
ad_Thickness	04	1104	1204	1304	1404	1504	1604	1704	1804	1904	2004	2104
ae_Thickness	05	1105	1205	1305	1405	1505	1605	1705	1805	1905	2005	2105
af_Thickness	06	1106	1206	1306	1406	1506	1606	1706	1806	1906	2006	2106
aa_Width	07	1107	1207	1307	1407	1507	1607	1707	1807	1907	2007	2107
ab_Width	08	1108	1208	1308	1408	1508	1608	1708	1808	1908	2008	2108
ac_Width	09	1109	1209	1309	1409	1509	1609	1709	1809	1909	2009	2109
ad_Width	10	1110	1210	1310	1410	1510	1610	1710	1810	1910	2010	2110
aa_Height	11	1111	1211	1311	1411	1511	1611	1711	1811	1911	2011	2111
ab_Height	12	1112	1212	1312	1412	1512	1612	1712	1812	1912	2012	2112
ac_Height	13	1113	1213	1313	1413	1513	1613	1713	1813	1913	2013	2113

data. The minimum acceptable confidence level was selected for the analysis. When the association rules were generated, the certainty of each of the rules was tested against the minimum acceptable confidence level. If the confidence of any rule was less than the defined level the rule was discarded otherwise it was kept in the final output.

In initial tests of the method, the association rule algorithm generated frequent itemsets and thousands of rules. Each rule had its confidence level, which was equal to or higher than the defined threshold value. When the rules were checked against the manufacturing process output some interesting facts were identified and these results are discussed in the next section. There is an important issue about the quality of the generated rules, and it is therefore very important to find the validity of each of the generated rules before time is spent analysing all the output, as this can be a very long, tedious, and time-consuming task. It is therefore advantageous to dispose quickly of any rules that are actually misleading or simply not valid.

4.3 Rule quality

Rules generated with very high support and confidence levels are less likely to be misleading than rules generated with lower support levels. This can be illustrated with the example data shown in Table 4. These data show 15 transactions or products, which have been carefully chosen for illustration purposes only.

There are many valid rules present in these data but two will be considered

1. 1612 → 1507
   (IF ‘ab’ section of ‘Height’ is nominal THEN ‘aa’ section of ‘Width’ is S-Upper.)
2. 1612 → 1303
   (IF ‘ab’ section of ‘Height’ is nominal THEN ‘ac’ section of ‘Thickness’ is H-Upper)

Table 4 Example data

ID	Data
1	1612, 1303, 1507, ...
2	1703, 1303, 1703, ...
3	1611, 1303, 1612, ...
4	1612, 1303, 1507, ...
5	1408, 1303, 1404, ...
6	2106, 1303, 1703, ...
7	1408, 1703, 1504, ...
8	1603, 1303, 1609, ...
9	1612, 1303, 1507, ...
10	1312, 1303, 1803, ...
11	1713, 1303, 1601, ...
12	1612, 1404, 1507, ...
13	1612, 1303, 1507, ...
14	1401, 2006, 1507, ...
15	1612, 1303, 1507, ...

where

nominal: 7.496–7.504
S-Upper: 7.504–7.508
H-Upper: 7.508–7.512

When the above rules are checked against the data, both seem to be valid, the first rule having 40 per cent support and 100 per cent confidence and the second rule also having support of 40 per cent and 83.3 per cent confidence. The first rule is a valid rule because 1507 and 1612 complement each other in the data, while the second rule is misleading as 1612 is complementing 1303 but 1303 does not complement 1612. Examination of the data shows that 1303 appears several times even when 1612 is not present so 1303 is independent of 1612. This situation clearly leads to the deduction that there is an association present in the data but its strength is uncertain. An association is strong only when both sides of the rule come together and neither element appears elsewhere in the data independently. The absence of one variable in the rule or too many appearances of one side of the rule makes it less important.

A popular technique for finding the correlation or the significance of the rules is the χ^2 test. χ^2 is popular for finding the correlation of bi-variant data in the statistics community. A higher value of χ^2 for the bi-variant (both ‘if’ statement and ‘then’ statement) shows a strong relationship and a lower value indicates a weak relationship. The strength of the relations can be found using the χ^2 table under the specific degree of freedom (1 in this case). In the above-quoted example the χ^2 value of the first rule was 11.42 and for the second rule it was only 0.5113, showing a confidence (strength) of more than 99 per cent in the first case and about 50 per cent in the second case. These results show that even though the confidence of the second rule was very high, when the quality of the rule is checked it resulted in a very poor rule, which is not significant.

5 RESULTS AND DISCUSSION

In order to include the dimensions, in the frequent itemsets or knowledge discovery process, that appear infrequently in the data, a very low support level was chosen for the original manufacturing data. Initially 2200 records were selected as training data sets, with only nine different attributes selected (chosen as priorities) and each of these had three sections, except one attribute which had only one section. This made a total of 25 × 11 unique items where 11 indicates the number of bands made between the upper and lower tolerance values of the dimensions at the different sections. The size of the most frequent itemset decreases as the support level increases and vice versa. Therefore in the first phase of this research, which involved finding the relationships between different dimensions of the output product, experiments were carried out running the association rule algorithm with 20 per cent support and 80 per cent confidence with each transaction having 25 items. The itemsets generated have a maximum of 10 items and there were 10 561 different combinations of frequent itemsets, each having between 2 and 10 elements. There were in the range of 20 000 significant rules generated. It would have been a very difficult task to analyse this many rules individually, therefore the χ^2 value of each of the rules was calculated. The χ^2 index for degree of freedom 1 and 95 per cent confidence in the strength of the relation of the rule is 3.841. Therefore any rules with index values less than 3.841 were discarded at this stage. This left 11 250 rules with a 95 per cent (or greater) confidence level in the relationship indicated by the rule.

There were still too many rules to analyse individually and therefore effort was initially concentrated on the rules that indicated one-to-one relationships. Efforts were then turned to rules that indicated one-to-two or two-to-one relationships and so on. By this stage, only 68 rules were found that indicated one-to-one relationships. One-to-one relationships are very important in identifying any design constraints of the product or any kind of manufacturing process limitation. Further examination of the remaining 68 rules indicated that there were 38 which did not provide any useful information. For example, within these 38 relationships were rules indicating that the 'nominal' value of one section corresponds to the 'nominal' value of another section. This is perfectly acceptable behaviour and indicates good manufacturing control for these dimensions and sections. However, the remaining 30 one-to-one rules revealed some very interesting and important facts. For example, some rules indicated that one particular 'nominal' dimension corresponds to another dimensional band being non-nominal. This type of result requires more attention, and provides valuable feedback for the design process, as such rules can help in redefining design constraints.

The extracted rules were then tested against new production data to see if the identified relationships are still valid in the new test data. The test results indicated that strong rules with high χ^2 values still hold in the new data with similar confidence of χ^2 index.

It is important to note that such relationships can indicate two important aspects of manufacturing. The first is where a design error may exist as the relationship shows that naturally the two correspondent dimensions do not have a 'nominal-to-nominal' relationship. The other important aspect could be identification of errors, faults, or limitations in the manufacturing process. Identification of these types of problems might need more careful reconsideration to improve manufacturing practice and strategies in order to remove any related faults. The remaining rules were all analysed and similar kinds of interesting design information have been found in them.

6 CONCLUSIONS AND FURTHER WORK

This paper presents a novel approach for the extraction of product design and manufacturing process improvement knowledge, using association rule techniques. Association rule is a data mining tool, which is commonly used for extracting relationships between different items or sale products in retail market areas. These techniques have not previously been applied and reported in manufacturing enterprise domains. The techniques reported here have been successful in substantial industrial case studies. This research shows that association rule techniques can be used on historical product data to extract information about the process limitations and knowledge of the interrelationships between particular product dimensions. Such information can then be fed back to establish design change requirements and product quality improvement constraints. The proposed technique is simple to use and can provide very valuable results.

As this methodology is very flexible and versatile, association rules could also be used on manufacturing process attribute data to discover relationships between different manufacturing processes and the output product dimensions from that process. In this way the ideal manufacturing process attributes can be determined for any specific output dimension and the best possible controlling parameter sets can therefore be discovered to achieve consistently high precision components or alternatively to identify production variables which lead to components going out of tolerance [**29**]. For example, in the case of fan blade manufacture, there are several processes involved – including diffusion bonding, superplastic forming, twisting, and machining. It is known that some dimensional parameters of the component are more greatly affected by particular processes than others. To determine the degree of the relationships, it is necessary to consider the key parameters for each process and the range of possible input values that these may take. Actual production data therefore need to be available for the key process parameters, so that they can be combined with the dimensional data for each blade, when it has passed through the process. Additional studies could also be made using CMMs or other measurement data from intermediate stages of its production as well from completed products. To apply the association rule methods, all dimensional data should be transformed as described in sections 3 and 4 of this paper, and in addition, suitable transformations need to be determined to translate the process parameter values into integer identifiers. If high-quality rules can be generated from the data (as in the research reported here), it should then be possible to use the knowledge identified to determine process constraints, limitations, and tolerances. The identification of precise knowledge of this type from actual production experiences, would enable resources to be targeted effectively for process improvements. Further research is currently being carried out in this area. The association rule methodology can also be used on historical operational databases to characterize process uncertainty and parameter estimation for better control of the manufacturing system by finding associations between the process variables and performance measurements [**30**].

It is also believed that the association rule technology described here has potential applications in the areas of predictive and preventative maintenance, and research and further case study experiments have recently started on data from historical calibration records and maintenance reports. In this way the technology might be used to determine very precise preventive maintenance schedules and numerous similar applications.

There are many issues still to be addressed in respect to the exploitation and reuse of both the knowledge discovered through data mining and data mining expertise that is gained through experimentation. Further work to be carried out in this research programme will be to identify ways to efficiently store mining information and the resulting rules in information and knowledge models so that they can be shared between different manufacturing set-ups across the manufacturing enterprise. Initial concepts for this research have been discussed by the current authors in reference [**31**].

REFERENCES

1 **Harding, J. A., Yu, B.,** and **Popplewell, K.** Information modelling: an integration of views of a manufacturing enterprise. *Int. J. Prod. Res.*, 1999, **37**(12), 2777–2792.

2 **Harding, J. A.** and **Yu, B.** Information-centred enterprise design supported by a factory data model and data warehousing. *Comput. Industry*, 1999, **40**, 23–36.

3 **Usama, M. F., Piatetsky-Shapiro, G., Smith, P.,** and **Uthurusamy, R.** Advances in knowledge discovery and data mining, 1996 (AAAI/MIT Press, Menlo Park, CA, USA).

4 **Netz, A.** Integration of data mining and relational databases. In Proceedings of the 26th International Conference on *Very large databases*, Cairo, Egypt, 2000.

5 **Maki, H.** and **Teranishi, Y.** Development of automated data mining system for quality control in manufacturing. *Lect. Notes Comput. Sci.*, 2001, **2114/2001**, 93–100.

6 **Braha, D.** *Data mining for design and manufacturing*, 2001 (Kluwer Academic Publishers, Norwell, MA, USA).

7 **Gertosio, C.** and **Dussauchoy, A.** Knowledge discovery from industrial databases. *J. Intell. Mfg*, 2004, **15**(1), 29–37.

8 **Maki, H., Maeda, A., Morita, T.,** and **Akimori, H.** Applying data mining to data analysis in manufacturing. International Conference on *Advances in production management systems*, Berlin, 1999, pp. 324–331 (Kluwer Academic Publishers).

9 **Shahbaz, M.** unpublished PhD Thesis, Loughborough University, Laghboragh, UK.

10 **Hashimot, K., Matsumoto, K.,** and **Terabe, M.** Applying data mining techniques to telecommunication services. *Jap. Soc. Artif. Intell.*, 2002, **17**(3), 320–325.

11 **Omer, F. A., Ertan, K.,** and **Piero, M.** Data Mining for Database Marketing at Garanti Bank. In *Data mining II* (Eds N. F. F. Ebecken and C. A. Brebbia), 2000 (WIT Press, UK).

12 **French, M. E.** *Form, structure and mechanism*, 1992 (McMillan, London).

13 **Pahl, G.** and **Beitz, W.** *Engineering design*, 1999 (Springer-Verlag, London).

14 **Hazelrigg, G. A.** *Systems engineering: an approach to information base design*, 1996 (Prentice-Hall Inc., New Jersey).

15 **Dym, C. L.** and **Levitt, R. E.** *Knowledge based systems in engineering*, 1991 (McGraw-Hill, New York).

16 **Romanowski, C. J.** and **Nagi, R.** A data mining-based engineering design support system: a research agenda. In *Data mining for design and manufacturing: methods and applications* (Ed. D. Braha), 2001, pp. 161–177 (Kluwer, London).

17 **Dym, C. L.** *Engineering design: a synthesis of views*, 1994, pp. 33–37 (Cambridge University Press, Cambridge, UK).

18 **Ertas, A.** and **Jones, J.** *The engineering design process*, 1996 (John Wiley, New York).

19 **Meneses, C.** and **Grinstein, G.** *Data mining*, 1998 (WIT Press/Computational Mechanics Publications, UK).

20 **SPSS.** *CRISP-DM 1.0 step-by-step data mining guide*, 1999 (SPSS Inc., Chicago, Illinois).

21 **Goebel, M.** and **Gruenwald, L.** A survey of data mining and knowledge discovery software tools. *SIGKDD Exploration*, 1999, **1**(1), 20–33.

22 **Agrawal, R., Imielinski, T.,** and **Swami. A.** Mining association rules between sets of items in large databases. In Proceedings of the ACM SIGMOD Conference, Washington DC, 1993.

23 **Agrawal, R.** and **Srikant, R.** Fast algorithms for mining association rules. In Proceedings of the VLDB Conference, Santiago, Chile, 1994.

24 **Brin, S., Motwani, R., Ullman, J. D.,** and **Tsur, S.** Dynamic itemset counting and implication rules for market basket data. In Proceedings of the ACM

SIGMOD International Conference on *Management of data*, Tucson, Arizona, 1997, pp. 255–264 (ACM Press, New York).

25 **Park, J. S., Chen, M. S.** and **Yu, P. S.** An effective hash based algorithm for mining association rules. In Proceedings of the ACM SIGMOD International Conference on *Management of data*, San Jose, California, 1995 (ACM Press, New York).

26 **Sarasere, A., Omiecinsky, E.,** and **Navathe, S.** An efficient algorithm for mining association rules in large databases. In Proceedings of the International Conference on *Very large databases*, Zurich, Switzerland, 1995.

27 **Toivonen, H.** Sampling large databases for association rules. *VLDB J.* 1996, 134–145.

28 **Dunham, M. H.** *Data mining, introductory and advanced topics*, 2003, p. 315 (Pearson Education Inc., New Jersey).

29 **Shahbaz, M., Srinivas,** and **Harding, J. A.** Knowledge extraction from manufacturing process and product databases using association rules. In *PDT Europe*, Stockholm, Sweden, 2004 (Eurostep AB, Stockholm Sweden).

30 **Srinivas, Harding, J.,** and **Shahbaz, M.** Agent oriented planning using data mined knowledge. In Proceedings of the 10th International Conference on *Concurrent engineering, adaptive engineering for sustainable value creation.* (Eds K.-D. Thoben, K. S. Pawar, and F. Weber) Seville, Spain, 2004, pp. 301–307 (Centre for Concurrent enterprising, University of Nottingham, UK).

31 **Shahbaz, M.** and **Harding, J. A.** An integrated data mining model for manufacturing enterprises. In Proceedings of the International Conference on *Manufacturing research*, 2003 Glasgow (Professional Engineering Publishing Limited, Bury St Edmonds and London).