APPLICATION OF DATA MINING TOOLS AND TECHNIQUES IN MATERIAL SELECTION

Pushpesh Pant and SriramPandey

Abstract - The purpose of this study is to investigate how data mining technique can be applied in materials informatics to extract knowledge from materials database.Studying material data sets from a data mining perspective can be usefulin material selection for design, manufacturing and other application of industrial engineering. This study employs predictive data mining technique to model a knowledge discovery system for the selection of materials that can satisfy the design specifications. Well known predictive method, Association Rule Mining (WEKA software)is used for this study. The algorithm of the association rule mining is implemented successively enabling it to unwrap the hidden pattern within the sample material database and the outcomes can be very useful for material engineers to speed up decision making process in manufacturing and other industrial engineering applications.

Index Terms - Association Rule Mining, Engineering Materials, Support, Confidence, WEKA software

IJSER

1. Introduction

There is a popular saying that "we are living in an information age" but we are actually living in a "data age". Terabytes or petabytes of data getting stored in our various storage devices, computer network, and the World Wide Web (WWW) every day from business, science, engineering, medicine, society and every other aspect of daily life.This furious growth in available data volume is a result of the computerization of our society. Brisk advances in data collection and storage technology have enabled the organization to gather vast amounts of data.However, extracting meaningful information out of large data sets has proven extremely challenging (Kamber *et al*., 2012; Tan, 2006).

Data mining is a technology that merges traditional data analysis methods with sophisticated algorithms for processing large volumes of data. It has also opened up exciting opportunities for exploring and analyzing new types of data and for old types of data in new ways.

Data mining is the process of extracting or mining useful information from large amounts of data. Data mining techniques are employed to a large database in order to find useful patterns that might otherwise remain known.They also provide capabilities to predict the outcome of a future observation, such as predicting whether the material satisfies the design specification or not (Kamber *et al*., 2012; Tan, 2006).

Among the many problems that are solved by data mining, two are of most importance. The first is 'unsupervised clustering' which is the task of finding subsets of the data such that the items from the same subset are most similar and items from the distinct subsets are more dissimilar. The second is (predictive modeling, supervised learning) where predictive mining tasks perform induction on the current data in order to make predictions (Witten *et al*., 2005).

1.1 Knowledge Discovery in Database

Data mining is a vital part of knowledge discovery in databases (KDD), which is a process of converting raw data into useful information, as shown in Figure 1.1.

Pushpesh Pant is currently pursing PhD in Industrial and System Engineering at IIT Kharagpur. He received his BE in Mechanical Engineering from SDM College of engineering and Technology, Dharwad, Karnataka, an M.Tech in Technology Management from Devi Ahilya University, Indore, Madhya Pradesh. His research focuses on industrial engineering, forecasting, material science, and data mining.

SriramPandey is currently pursuing M.Tech. from IIT Kharagpur. He received his BE in Mechanical Engineering from SDM College of engineering and Technology, Dharwad, Karnataka. His research focuses on material science, fluid mechanics, ocean engineering etc.

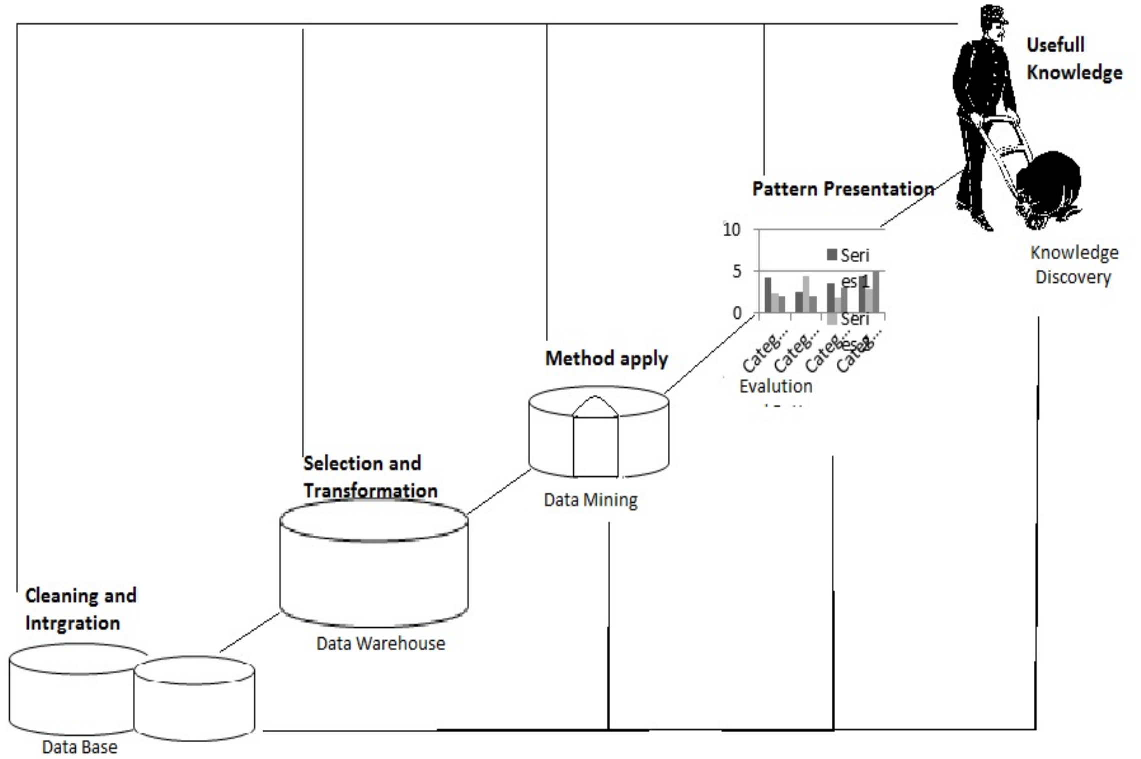


Fig 1.1 Knowledge discovery process (Pant et al., 2016).

IJSER

This process consists of series of transformation steps, from data pre-processing to post-processing of data mining results, as shown in Figure 1.2 (Tan, 2006). The input data can be stored in many formats (flat files, spreadsheets, or relational tables). The main function of data pre-processing is to convert the raw input data into a suitable format further analysis. The steps involved in data pre-processing are blending of data from various sources, data cleaning by removing noise and duplicate observations and selecting records which are relevant to the data mining task at hand. Data pre-processing is perhaps the most challenging and time-consuming step in the overall knowledge discovery process.

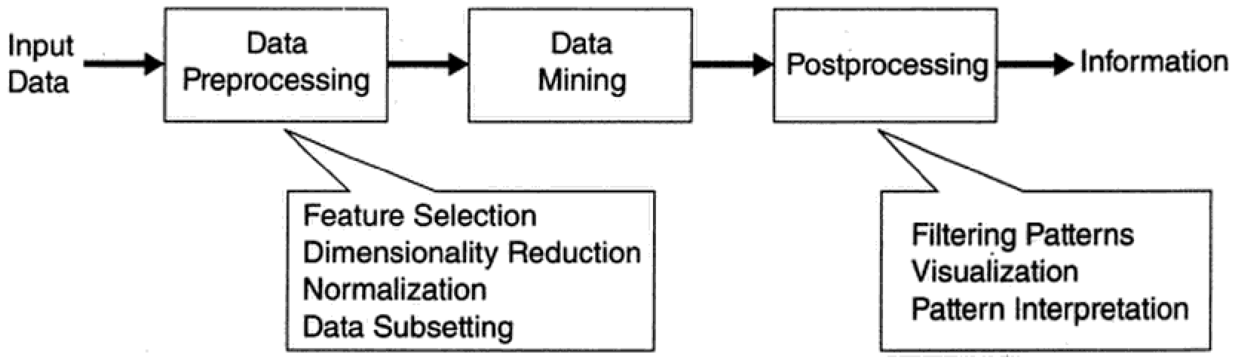


Fig 1.2 The process of knowledge discovery in databases (Tan, 2006).

Post-processing involves filtering, visualization and interpretation of pattern from the mined data. It also ensures that only valid and useful results are incorporated into a decision support system.

1.2 Material Database Management

During the rapid development of materials sciences, researchers are facing more new issues, and a lot of traditional analytic techniques are not effective any longer under these situations. Firstly, materials is becoming much more complicated, engineering-driven, and multifunctional (Hunt, 2006). Secondly, investigating the relations between structure, properties is an important research object in materials science, however, these relations are always non-linear, and thus it is difficult to model them and seek patterns. Furthermore, materials databases are becoming increasingly complicated, and the quantity increases in geometric series. In recent years, researchers have begun to study the combined effect of data mining and material science. The material informatics are emerging in material science as a new research topic, and has already changed our experimental methods and way of thinking in materials research, and will lead even more changes (Song, 2004; Suh & Rajan, 2009).

Data mining is an interdisciplinary field merging ideas statistic, artificial intelligence, machine learning, databases, and parallel and distributed computing, and is a very useful tool to integrate information and theory for materials discovery.

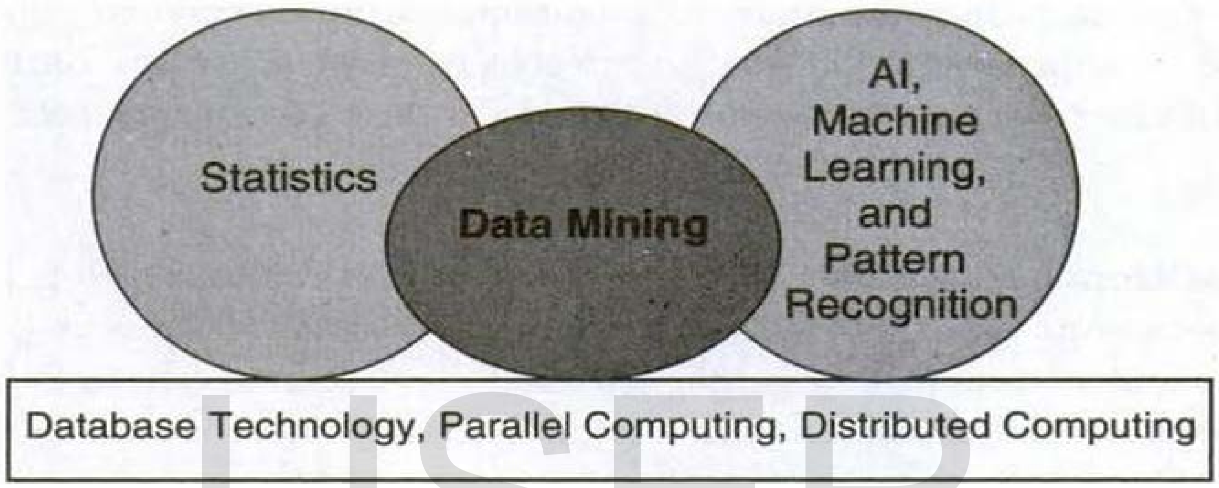

Statistics
Data Mining
AI, Machine Learning, and Pattern Recognition
Database Technology, Parallel Computing, Distributed Computing

Fig 1.3 Data mining as a conflux of many disciplines (Tan, 2006).

IJSER

Therefore, data mining can be used to extract knowledge and insight from massive materials databases. In practice, with the help of data mining, accurate and fast prediction of crystal structure, material properties and behaviors based on the materials data has been successfully implemented in many tasks. In addition, combinatorial materials and new data in large quantities generated by high-throughput experiments can provide a wider platform for data mining to processing those data (Rodger &Cebon, 2006).

Quantitative computational models and improved algorithms, higher computer performance and professional software make data mining methods more easily and more effective. Data mining has two primary functions: pattern recognition and prediction, both of them are fundamentals to understand materials behaviors. So, In order to perform these tasks, it is necessary to have a dataset containing useful information which can be helpful while predicting materials and their applications. In material informatics, research areas are mainly focused on following tasks (Hunt, 2006; Rodger *et a*l., 2006).

1) *Data standards:* There are thousands of materials databases in different formats (Westbrook, 2003), and they are difficult to communicate with each other. To standard these databases and to integrate materials data into a single database is the first important task of materials informatics (Song, 2004) to enable knowledge discovery.
2) *Data mining on materials data:* There is an enormous range of possible new materials, and it is often difficult to physically model the relationships between constituents, and processing, and final properties. Data mining has the abilities to search and analyze material data and find potential, previously unknown patterns rules. It involves selecting, exploring and modeling large amounts of data to uncover previously unknown patterns from large databases (Rodgers, 2003; Chen *et al*., 2009). Data mining involves some high-effective computational algorithms, such as neural networks, genetic algorithm, etc.
3) *Cluster analysis:* As an exploratory data analysis tool, can sort different materials or properties into groups

in such a way that the degree of association between two objects is maximal if they belong to the same group and minimal otherwise. And, cluster analysis can be integrated with high-throughput experimentation for rapidly screening combinatorial data.

4) *Anomaly detection:* In properties analysis or combinatorial experiments, anomaly detection is used to identify anomalies, especially to assess the uncertainty and accuracy of results and distinguish between true discoveries and false-positive results.
5) *Association Analysis:* Association rule mining is good at finding patterns, and can be used to develop heuristic rules for materials behavior based on large data sets.
6) *Predict Modeling:* Some machine learning algorithms can be used for material class prediction, and materials classification models such as Neural Network (NN) can be built up predict models.
7) *Outlier Analysis:* In properties analysis or combinatorial, outlier analysis is used to identify peculiarities, especially to compute the uncertainty and accuracy of results, and distinguish between true discoveries and false positive results.
8) *Material Visualization:* Reconstruction of material structure information based on materials data would help researchers to analyze the relationships between material structure and material properties.
9) *Organization and management of material data*: In order to meet materials researchers' different needs, satisfy the need of research and production, to construct the materials data into a whole and single coherent database, efficient Materials Database Management Systems(MDBMS)is very necessary (Hrubiak *et al*., 2009)

1.3 WEKA (Waikato Environment for Knowledge Analysis)

The Waikato Environment for Knowledge Analysis (WEKA) came into the picture in need for a workbench that would allow researchers and practioners, aeasy access to state-of-the-art techniques in machine learning. It was envisioned that WEKA would not only provide a toolbox of learning algorithms, but also a framework inside which researchers could implement new algorithms without having to be concerned with supporting infrastructure for data manipulation and scheme evaluation. Nowadays, WEKA is recognized as a benchmark system in data mining and machine learning (Shapiro, 2005). It has achieved widespread acceptance within academia and business circles and has become a widely used tool for data mining research(Witten *et al*., 2005).Giving users free access to the source code has enabled a thriving community to develop and facilitated the creation of many projects that incorporate or extend WEKA.

The WEKA workbench aims to provide a comprehensive collection of machine learning algorithms and data preprocessing tools to researchers and practitioners alike. It allows users to quickly try out and compare different machine learning methods on new data sets. Its modular, extensible architecture allows sophisticated data mining processes to be built up from the wide collection of base learning algorithms and tools provided. The workbench includes algorithms for regression, classification, clustering, association rule mining and attribute selection. Preliminary exploration of data is well catered for by data visualization facilities and many preprocessing tools (shearer, 2000).

WEKA has several graphical user interfaces that enable easy access to the underlying functionality. The main graphical user interface is the "Explorer". It has a panel-based interface, where different panels correspond to different data mining tasks. In the first panel, called "Preprocess" panel, data can be loaded and transformed using WEKA's data preprocessing tools, called "filters". Data can be loaded from various sources, including files, URLs, and databases. Supported file formats include WEKA's own ARFF format, CSV, LibSVM's format, and C4.5's format. It is also possible to generate data using an artificial data source and edit data manually using a dataset editor.

The second panel in the Explorer gives access to WEKA's classification and regression algorithms. The corresponding panel is called "Classify" because regression techniques are viewed as predictors of "continuous classes". By default, the panel runs a cross-validation for a selected learning algorithm on the dataset that has been prepared in the Preprocess panel to estimate predictive performance. It also shows a textual representation of the model built from the full data set. However, other modes of evaluation, e.g. based on a separate test set, are also supported. If applicable, the panel also provides access to graphical representations of models, e.g. decision trees. Moreover, it can visualize prediction errors in scatter plots, and also allows evaluation via ROC curves and other "threshold curves". Models can also be saved and loaded in this panel.

Along with supervised algorithms, WEKA also supports the application of unsupervised algorithms, namely

IJSER

clustering algorithms and methods for association rule mining. These are accessible in the Explorer via the third and fourth panel respectively. The "Cluster" panel enables users to run a clustering algorithm on the data loaded in the Preprocess panel. It provides simple statistics for evaluation of clustering performance: likelihood-based performance for statistical clustering algorithms and comparison to "true" cluster membership if this is specified in one of the attributes in the data. If applicable, visualization of the clustering structure is also possible, and models can be stored persistently if necessary. WEKA's support for clustering tasks is not as extensive as its support for classification and regression, but it has more techniques for clustering than for association rule mining, which has up to this point been somewhat neglected(Shapiro, 2005).

2. Literature review

This section explains the contributions of researchers and practitioners in different areas of design, manufacturing and other application of industrial engineering. The literature was searched thoroughly from different journals, personal web pages, the internet, and websites. This review is particularly focused on data-mining applications and case studies in material selection for manufacturing and design processes and closely related fields.

Engineering Design

It is a multidisciplinary, multidimensional, and non-linear decision-making process where parameters, actions, and components are selected. This selection is mainly based on historical data, information, and knowledge. It is, therefore,an important area for data mining applications and even though only a few papers have reported applications of data mining in engineering design, this has been an area of increased research interests in recent years. The importance of considering how a product should be manufactured during the design stage and the constraints imposed on a design by particular material, manufacturing processes and technologies have been accepted for many years. There is indeed a great potential for data mined knowledge to integrate material selection, manufacturing, product characteristics, and the engineering design processes.

Sim and Chan (1992) developed a knowledge-based system for the selection of rolling bearings. They used heuristic knowledge supported by a manufacturer's catalogue to optimize design specifications by matching the temporal data of the new product against the knowledge base. (Kusiak et al., 2000) proposed a rough-set theory approach to predict product cost. Ishino and Jin (2001) used data mining for knowledge acquisition in design from the data obtained through observing design activities using a CAD system. They developed a method called Extended Dynamic Programming to extract the knowledge. (Romanowski and Nagi (2001) proposed a design system which supports the feedback of data mined knowledge from the life cycle data to the initial stages of the design process. Giess*et al*. (2002, 2003) mined a manufacturing and assembly database of gas turbine rotors to determine and quantify relationships between the various balance and vibration tests and highlight critical areas. This knowledge could then be fed back to the designers to improve tolerance decisions in the future design of components. They used a decision tree at the initial stage to determine appropriate areas of investigation and to identify problems with the data. At the next stage, a neural network was used to model the data. Hamburg (2002) applied data mining techniques to support product development by analyzing global environment aspects, market situation, strategy, philosophy, and culture of the manufacturing and customer behavior. He utilized a decision-tree algorithm to determine and integrate the enterprise data in the product development. Romanowski and Nagi (2004, 2005) applied a data-mining approach for forming generic bills of materials (GBOMS), entities that represent the different variants in a product family and facilitate the search for similar designs and the configuration of new variants. By combining data-mining approaches such as text and tree mining in a new tree union procedure that embody the GBOM and design constraints in constrained XML, the technical difficulties associated with a GBOM are resolved.

Kim and Ding (2005) presented a data mining aided optimal design method capable of finding a competitive design solution with a relatively low computation cost. They applied the method to facilitate the optimal design of fixture layout in a four-station SUV side panel assembly process.

Manufacturing Systems

IJSER

Data collection in manufacturing is common but its use tends to be limited to rather few applications. Machine learning and computational intelligence tools provide excellent potential for better control of manufacturing systems, especially in complex manufacturing environments where detection of the causes of problems is difficult. Piatesky-Shapiro *et al*. (1999) argued that the data mining industry is coming of age. However, this review of data mining in manufacturing shows that although there are several areas in manufacturing enterprises that have benefited from data-mining algorithms, there are still numerous areas that could benefit further (Harding *et al*., 2003). In manufacturing environments, the need and importance of data collection is ever present for statistical process control purposes. Lee (1993) discussed and suggested several principles leading to a knowledge-based factory environment utilizing the data collected over several stages of the manufacturing-related processes. A comparative study of implicit and explicit methods to predict the non-linear behavior of the manufacturing process, using statistical and artificial intelligence tools, was discussed by (Kim and Lee, 1997).

Semiconductor manufacturing is complex and faces several challenges relating to product quality, scheduling, work in process, cost reduction, and fault diagnosis. To overcome these problems several methods and systems have been developed, e.g., Rule-Based Decision Support Systems (RBDSS) (Adachi *et al*., 1989), CAQ (Whitehall *et al*., 1990), Knowledge Acquisition from Response Surface Methodology (KARSM), and GID3 (Irani *et al*., 1993) or generalized ID3, a decision-tree algorithm for fault diagnostics and decision making have been developed and used. Gardener and Bieker (2000) showed substantial savings in the manufacture of semiconductors by applying decision tree algorithms and neural networks to solve the yield problem in the wafer manufacture. Sebzalli and Wang (2001) applied principal component analysis and fuzzy c-means clustering to a refinery catalytic process to identify operational spaces and develop operational strategies for the manufacture of desired products and to minimize the loss of product during system changeover. Four operational zones were discovered, with three for product grade and the fourth region giving a high probability of producing the off-specification product. (Lee and Park, 2001) used self-organizing maps and Last and Kandel (2001) applied information fuzzy networks for quality checks and extracted useful rules from their model to check the quality of the products. Kusiak (2001) proposed a rule-structuring algorithm that can handle data from different sources to extract rules, which is very helpful in semiconductor manufacturing. The algorithm formed relevant meta-structures enhancing the utility of the extracted knowledge. Dabbas and Chen (2001) proposed the consolidation and integration of data from the different semiconductor manufacturing sources into one database to generate different factory performance reports. Their method can be further exploited to use data mining to extract information from these reports.

Different data mining tools for improvement in integrated circuit manufacturing were presented in McDonald (1999). Another successful application of a sophisticated data mining algorithm was reported by Fountain *et al*. (2003). They used the Naïve Bayes probabilistic model in their theoretic decision-making approach to optimize testing of dies (ICs in the wafer form) during a die-level functional test. Their results showed substantial savings in testing costs and hence reduced overall costs compared with other testing policies, such as “exhaustive,” “package all,” and “Oracle.”

An interesting area of research in manufacturing enterprises has been a determination of optimal machining parameters to minimize machining errors such as tool wear, tool breakage, and tool deflection, which could result in slower production rates and increased costs. Park and Kim (1998) reviewed different techniques based on CAD systems, operational research, and computational intelligence to determine the optimal solutions to these errors and for online adaptive control using knowledge-based expert systems. Other knowledge-based systems have also been proposed (Liao*et al*., 1999) for condition interpretation of tools and quality of the products.

Performance and quality issues have also been considered while applying data mining techniques in manufacturing process related areas. Gertosio and Dussauchoy (2004) have used linear regression analysis to determine and establish the relationships between test parameters and the performance of truck engines. Their simple methodology showed up to 25% savings in test process time. A method to reduce the component testing time required before the assembly was proposed by (Yin *et al*., 2001). They applied genetic and rough-set algorithms on past test data to find the optimal test criteria to substantially reduce the overall testing time. Another successful application of a regression model is presented (Feng and Wang, 2004) to predict the performance of the knurling process and the quality of the knurls. Similar results were achieved by using both regression and neural networks.

Efforts have also been made to develop models to study the entire factory or enterprise data altogether to discover the problem areas instantly affecting any subsequent processes. Maki *et al*. (1999, 2001) developed an intelligent system in Hitachi for online data analysis using a data mining approach. Their system used a rule induction algorithm to extract rules using an automated data mining engine and delivered the results using an intranet for easy access. Adams (2002) analyzed the different software that can be used to mine a factory's data and compare the features of information sharing. Shahbaz's (2003) integrated data mining model is the next level of knowledge sharing, as once the data are mined the relevant data and data mining results can be shared within the factory and beyond at other sites, by using a neutral data format. Shahbaz*et al*. (2004) used association rules for product design improvement and applied supervised association rules for controlling the product dimensions by controlling the process variables using supervised association rules (Shahbaz*et al*., 2004; 2005). Their methodology can be used as an alternative and/or a support to the design of experiment methodology.

Chen *et al*. (2006) applied data mining in hyperspace to identify material properties. They used MasterMiner to build a hyperspace data mining model, which uses non principle factors or most relevant variables and built a mathematical model to find the solution equation in non dimensional space for a specific material property. This technique is useful in chemical and material industry where different variables affect one or more properties of the material or chemical reaction. Interesting research has also been done by (Mere *et al*., 2004) to determine the optimal mechanical properties of galvanized steel by using a combination of clustering and neural networks. Clustering was used in the first instance and then neural networks were applied to the clusters to predict the mechanical properties of the steel.

3. Problem Definition

Data mining is widely accepted tool in the area of statistics, machine learning, artificial intelligence, and material science because of its ability to extract meaningful pattern and information out of large database. Although it is not often to integrate data mining and material database, few papers have reported applications of data mining in material selection using a Naïve Bayesian Classifier algorithm, Rule-Based Classifier, Decision Tree Classifier. Most of the work we have seen in the literature is based on either finding pattern in the material's crystal structure or material chemistry. But there is not much study focusing on how to select a particular material out of the large material database for specific application.

For example, manufacturing of car chassis requires a material which should have high strength; high toughness and light weight and aluminum alloy and steel alloy satisfy these properties. The Industry is using these alloys for the manufacturing of chassis since the inception of motor car. So the challenge is to find another material which is more economical and effective from the earlier. And data mining would play a big role in finding alternate materials from the material database having a hundred thousand of materials.

The objective of this study are as follows:

- ➢ To investigate the data mining techniques for material selection that can be beneficial for design, manufacturing and other application of industrial engineering.
- ➢ To discuss how data mining can be applied in materials informatics by using materials data.

4. Material Database

In the recent era, science and technology is advancing rapidly. Accordingly, material science and engineering change with each passing day, and new ideas, methods, techniques and materials appear one after another. This leads to the huge amount of materials database which cannot be tackled manually, so data mining is becoming one of the general areas of material selection and development (Callister, 2007; Askeland, 2012).

TABLE 4: Steel Alloys with their Mechanical Properties

	Steel Alloy	Mechanical Properties

IJSER

1.	Ferrium M54	High Strength, Toughness, Corrosion Resistant
2.	Crucible Steel	Hardness, Toughness, Abrasion Resistant
3.	Hadfield Steel	Toughness, Abrasion Resistant, High Impact Strength
4.	High Speed Steel	Toughness, Abrasion Resistant, Hot Hardness, High Bending Strength, High Thermal Conductivity
5.	Maraging Steel	High Strength, Toughness, Corrosion Resistant
6.	Reynolds 531	Toughness, Light Weight, High Strength
7.	Reynolds 525	Toughness, Light Weight, High Strength
8.	Reynolds 520	Toughness, Light Weight, High Strength
9.	Tool Steel	Hot Hardness, Toughness, Abrasion Resistant, High Thermal Conductivity
10.	Weathering Steel	High Strength, Corrosion Resistant
11.	Wootz Steel	Toughness, High Impact Hardness, Plastic Properties
12.	Al-6XN	Corrosion Resistant, Weldable, High Strength, Formability
13.	Celestrium	High Resistant, Corrosion Resistant
14.	Marine Grade Steel	Corrosion Resistant, Heat Resistant, Weldable, High Fatigue Strength
15.	Alloy 28	Weldable, Formability, Corrosion Resistant, Abrasion Resistant
16.	Surgical Stainless Steel	High Strength, Formability, Corrosion Resistant
17.	Zeron 100	High Strength, Corrosion Resistant, Pitting Resistant
18.	Ferrium S53	High Strength, Corrosion Resistant
19.	Ferrium C61	High Strength, Toughness, Hardness

IJSER

19 steel alloys are considered for this study to show how a material database can give relevant information using data mining algorithm. Similarily, it is possible with hundred thousands of alloys.

5. Proposed Technique: Association Rule Mining

Prediction is one of the core tasks in data mining. In the above materials database, we use *Association Rule Mining* to predict the material for specific applications.

5.1 Association Rule Mining

It is an "If-Then" relationship. If it happens, what is most likely to happen next? It is a popular and well researched method for discovering interesting relation between variables in large databases. Association rule show attributes value conditions that occur frequently together in a given dataset (Kamber et al., 2011).

$$\text{Support} = \{(X\ U\ Y).\ \text{count}\} / (n) \tag{1}$$

The Model: Data

- $I = \{ i_1, i_2, i_3, - - - - i_0 \}$: A set of all the items
- Transaction, T: A set of items such that $T \subseteq I$
- Transaction database, D : A set of transactions, $T = \{t_1, t_2, \ldots, t_n\}$

The Model: Rules

- A transaction $T \subseteq I$ contains a set $X \subseteq I$ of some items, if $X \subseteq T$
- An association rule is an implication of the form $X \rightarrow Y$, where $X,Y \subseteq I$ and $X \cap Y = \varnothing$

Rule Strength Measure:

- Support is a rule to find the value of X with respect to T which is defined as the proportion of transaction in the database that contains the item-set X. Support is an indication of how frequently the items appear in the database.

- Confidence gives a value of the rule, X→Y, with respect to a set of transaction that contains X which also contains Y. Confidence indicates the number of times the if/then statements have been found to be true.

$$\text{Confidence} = \{(X \cup Y).\text{ count}\} / (X.\text{ count}) \quad (2)$$

5.2 Data Mining Algorithm

The Algorithm:

- Iterative algorithm (also called level-wise search): Find all 1-item frequent itemsets; then all 2-item frequent itemsets, and so on.
  In each iteration k, only consider itemsets that contain some k-1 frequent itemset.
- Find frequent itemsets of size 1: F_1
- From $k = 2$

C_k = candidates of size k: those itemsets of size k that could be frequent, given F_{k-1}

F_k = those itemsets that are actually frequent, $F_k \subseteq C_k$(need to scan the database once).

Ordering of Items:

- The items in I are sorted in lexicographic order (which is a total order).
- The order is used throughout the algorithm in each itemset.
- $\{w[1], w[2], \ldots, w[k]\}$ represents a k-itemset$w$ consisting of items $w[1], w[2], \ldots, w[k]$, where $w[1] < w[2] < \ldots < w[k]$ according to the total order.

Detailed Algorithm:

Algorithm Apriori(T)

```
    $C_1 \leftarrow$init-pass($T$);
    $F_1 \leftarrow \{f \mid f \in C_1, f.\text{count}/n \geq minsup\}$;     // n: no. of transactions in T
    for ($k = 2$; $F_{k-1} \neq \emptyset$; $k$++) do
        $C_k \leftarrow$ candidate-gen($F_{k-1}$);
        for each transaction $t \in T$do
        for each candidate $c \in C_k$do
            if$c$ is contained in $t$then
            $c.count$++;
        end
        $F_k \leftarrow \{c \in C_k \mid c.count/n \geq minsup\}$
        end
return$F \leftarrow \bigcup_k F_k$;
```

Apriori Candidate Generation:

- The candidate-gen function takes F_{k-1} and returns a superset (called the candidates) of the set of all frequent k-itemsets. It has two steps.
  join step: Generate all possible candidate itemsetsC_k of length k

prune step: Remove those candidates in C_k that cannot be frequent.

- Candidate Generation Function

```
Function candidate-gen($F_{k-1}$)
    $C_k \leftarrow \emptyset$;
    forall$f_1, f_2 \in F_{k-1}$
        with$f_1 = \{i_1, \ldots, i_{k-2}, i_{k-1}\}$
        and$f_2 = \{i_1, \ldots, i_{k-2}, i'_{k-1}\}$
        and$i_{k-1} < i'_{k-1}$do
    $c \leftarrow \{i_1, \ldots, i_{k-1}, i'_{k-1}\}$;            // join $f_1$ and $f_2$
    $C_k \leftarrow C_k \cup \{c\}$;
    foreach ($k$-1)-subset $s$ of $c$do
        if($s \notin F_{k-1}$) then
        delete$c$ from $C_k$;          // prune
    end
    end
    return$C_k$;
```

6. Results and Discussions

In this study, association rule mining is applied to the material data set to reveal the need of particular material for specific applications. The best known constraints are minimum thresholds on support and confidence. The support (X) of an item set X is defined as the proportion of transactions in the data set which contain the item set. The confidence of a rule is defined as:

$$\text{Confidence } (X \rightarrow Y) = \text{support } (XUY) / \text{support}(X) \quad (3)$$

IJSER

6.1. Manual Calculation for Few Pair of Itemsets

TABLE 6: Support and confidence for each pair of transaction

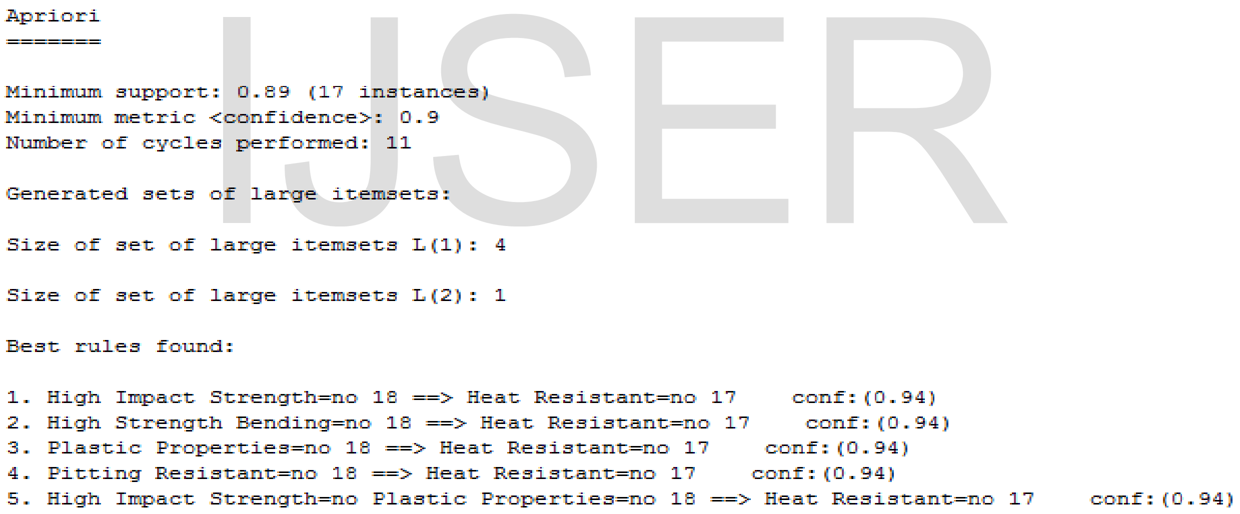
X → Y (Transaction)	Support	Confidence
Toughness → High Strength	42%	73%
High Strength → Toughness	42%	57%
Corrosion Resistant → High Strength	47%	90%
Light Weight → Toughness	16%	100%
Light Weight → High Strength	16%	100%
Hardness → Toughness	26%	100%

This is the manual method of conducting association rule mining just to show how it actually works. In the above table, we calculated support for different transaction (transaction between mechanical properties) to find the value of X with respect to T which is defined as the proportion of transaction in the database that contains the item-set X and confidence gives a value of the rule, X→Y, with respect to a set of transaction that contains X which also contain Y

6.2 Output in WEKA Software

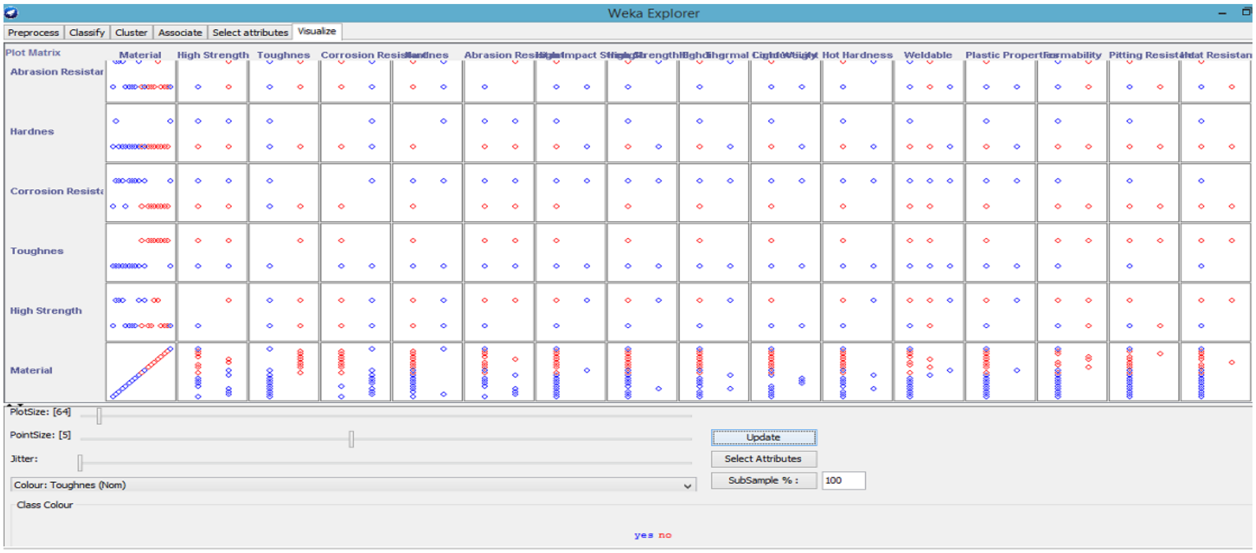
Apriori is a method of association rule mining. In conducting apriori in WEKA, we get the following result which is shown in the figure below. This result is based on our material database (containing 19 steel alloys).

IJSER



```
Apriori
=======

Minimum support: 0.89 (17 instances)
Minimum metric <confidence>: 0.9
Number of cycles performed: 11

Generated sets of large itemsets:

Size of set of large itemsets L(1): 4

Size of set of large itemsets L(2): 1

Best rules found:

1. High Impact Strength=no 18 ==> Heat Resistant=no 17    conf:(0.94)
2. High Strength Bending=no 18 ==> Heat Resistant=no 17    conf:(0.94)
3. Plastic Properties=no 18 ==> Heat Resistant=no 17    conf:(0.94)
4. Pitting Resistant=no 18 ==> Heat Resistant=no 17    conf:(0.94)
5. High Impact Strength=no Plastic Properties=no 18 ==> Heat Resistant=no 17    conf:(0.94)
```

Fig 6.1 Material selection rule in WEKA.



PlotSize: [64]

PointSize: [5]

Jitter:

Colour: Toughnes (Nom)

SubSample % : 100

Fig 6.2 Materials associated properties visualization.

The above figure shows the graphical representation between the mechanical properties where blue color indicates positive association and red color indicates negative association.

7. Conclusion

The primary aim of this study was to show how a simple data mining technique can be applied to answer a few specific questions on materials. Generally it is very difficult to choose a particular material for specific application manually because of very large and complex database available to us. This paper focused on how data mining technique such as association rule would be helpful in predicting suitable material for different applications. By applying data mining to material database we have got the following results,

- ➢ Integration of association rule mining with a material database can be helpful in selecting the right type of material for a specific application.
- ➢ It will minimize material selection time.
- ➢ It provides a way to switch from conventionally used material for a particular application with the new one.

References

[1] Adachi, T., Talavage, J. J., &Moodie, C. L. (1989). A rule-based control method for a multi-loop production system. *Artificial Intelligence in Engineering*, *4*(3), 115-125.

[2] C. Shearer. The CRISP-DM model: The new blueprint for data mining. Journal of Data Warehousing, 5(4), 2000.

[3] Chen, N., Zhu, D. D., & Wang, W. (2000). Intelligent materials processing by hyperspace data mining. *Engineering Applications of Artificial Intelligence*, *13*(5), 527-532.

[4] Dabbas, R. M., & Chen, H. N. (2001). Mining semiconductor manufacturing data for productivity improvement—an integrated relational database approach. *Computers in Industry*, *45*(1), 29-44.

[5] Feng, C. X. J., & Wang, X. F. (2004). Data mining techniques applied to predictive modeling of the knurling process. *Iie Transactions*, *36*(3), 253-263.

[6] Fountain, T., Dietterich, T., &Sudyka, B. (2003, January). Data mining for manufacturing control: an application in optimizing IC tests. In *Exploring artificial intelligence in the new millennium* (pp. 381-400). Morgan Kaufmann Publishers Inc.

[7] G. Piatetsky-Shapiro. KDnuggets news on SIGKDD service award.http://www.kdnuggets.com/news/2005/n13/2i.html, 2005.

[8] Gardner, M., & Bieker, J. (2000, August). Data mining solves tough semiconductor manufacturing problems. In *Proceedings of the sixth ACM SIGKDD international conference on Knowledge discovery and data mining*(pp. 376-383). ACM.

[9] Gertosio, C., &Dussauchoy, A. (2004). Knowledge discovery from industrial databases. *Journal of Intelligent Manufacturing*, *15*(1), 29-37.

[10] Giess, M. D., Culley, S. J., & Shepherd, A. (2002, January). Informing design using data mining methods. In *ASME 2002 International Design Engineering Technical Conferences and Computers and Information in Engineering Conference* (pp. 207-215). American Society of Mechanical Engineers.

[11] Giess, M. D., &Culley, S. J. (2003). Investigating manufacturing data for use within design. In *DS 31: Proceedings of ICED 03, the 14th International Conference on Engineering Design, Stockholm*.

IJSER

[12] Hamburg, I. (2002). Improving Computer Supported Environmental Friendly Product Development by Analysis of Data.

[13] Han, J., Pei, J., & Kamber, M. (2011). *Data mining: concepts and techniques*. Elsevier.

[14] Harding, J. A., Shahbaz, M., &Kusiak, A. (2006). Data mining in manufacturing: a review. *Journal of Manufacturing Science and Engineering*, *128*(4), 969-976.

[15] Harding, J. A., Shahbaz, M., &Kusiak, A. (2006). Data mining in manufacturing: a review. *Journal of Manufacturing Science and Engineering*, *128*(4), 969-976.

[16] Hedau, V., Pant, P., & Sharma, K. (2016). Material Selection using Association Rule Mining. *International Journal of Advanced Research in Computer Science*, *7*(3).

[17] Hrubiak, R., George, L., Saxena, S. K., & Rajan, K. (2009). A materials database for exploring material properties. *JOM*, *61*(1), 59-62.

[18] Hunt Jr, W. H. (2006). Materials informatics: Growing from the bio world. *JOM*, *58*(7), 88-88.

[19] Irani, K. B., Cheng, J., Fayyad, U. M., &Qian, Z. (1993). Applying machine learning to semiconductor manufacturing. *iEEE Expert*, *8*(1), 41-47.

[20] Ishino, Y., & Jin, Y. (2001). Data mining for knowledge acquisition in engineering design. In *Data mining for design and manufacturing* (pp. 145-160). Springer US.

[21] J. R. Rodgers, "Materials informatics: Knowledge acquisition for materials design.," ABSTRACTS OF PAPERS OF THE AMERICAN CHEMICAL SOCIETY, vol. 226, p. U302-U303, 2003.

[22] Kim, P., & Ding, Y. (2005). Optimal engineering system design guided by data-mining methods. *Technometrics*, *47*(3), 336-348.

[23] Kim, S. H., & Lee, C. M. (1997). Nonlinear prediction of manufacturing systems through explicit and implicit data mining. *Computers & industrial engineering*, *33*(3), 461-464.

[24] Kusiak, A. (2001). Rough set theory: a data mining tool for semiconductor manufacturing. *IEEE transactions on electronics packaging manufacturing*, *24*(1), 44-50.

[25] Kusiak, A., Kernstine, K. H., Kern, J. A., McLaughlin, K. A., & Tseng, T. L. (2000, May). Data mining: medical and engineering case studies. In *Industrial Engineering Research Conference* (pp. 1-7).

[26] Last, M., &Kandel, A. (2001). Data mining for process and quality control in the semiconductor industry. In *Data mining for design and manufacturing*(pp. 207-234). Springer US.

[27] Lee, J. H., & Park, S. C. (2001). Data mining for high quality and quick response manufacturing. In *Data mining for design and manufacturing* (pp. 179-205). Springer US.

[28] Lee, M. H. (1993). The knowledge-based factory. *Artificial intelligence in Engineering*, *8*(2), 109-125.

[29] Lian-Yin, Z., Li-Pheng, K., &Sai-Cheong, F. (2001). Derivation of decision rules for the evaluation of product performance using genetic algorithms and rough set theory. In *Data mining for design and manufacturing* (pp. 337-353). Springer US.

[30] Maki, H., &Teranishi, Y. (2001, September). Development of automated data mining system for quality control in manufacturing. In *International Conference on Data Warehousing and Knowledge Discovery* (pp. 93-100). Springer Berlin Heidelberg.

[31] Maki, H., Maeda, A., Morita, T., &Akimori, H. (1999). Applying data mining to data analysis in manufacturing. In *Global Production Management* (pp. 324-331). Springer US.

[32] McDonald, C. J. (1999). New tools for yield improvement in integrated circuit manufacturing: can they be applied to reliability?. *Microelectronics Reliability*, *39*(6), 731-739.

[33] OrdieresMeré, J. B., González Marcos, A., González, J. A., &Lobato Rubio, V. (2004). Estimation of mechanical properties of steel strip in hot dip galvanising lines. *Ironmaking& steelmaking*, *31*(1), 43-50.

[34] Park, K. S., & Kim, S. H. (1998). Artificial intelligence approaches to determination of CNC machining parameters in manufacturing: a review. *Artificial Intelligence in Engineering*, *12*(1), 127-134.

[35] Piatetsky-Shapiro, G. (1999). The data-mining industry coming of age. *IEEE Intelligent Systems*, *14*(6), 32-34.

[36] Rodgers, J. R., & Cebon, D. (2006). Materials informatics. *MRS bulletin*, *31*(12), 975-980.

[37] Romanowski, C. J., &Nagi, R. (2001). A data mining for knowledge acquisition in engineering design. *Data mining for design and manufacture: Methods and applications*, 161-178.

[38] Romanowski, C. J., &Nagi, R. (2005). On comparing bills of materials: a similarity/distance measure for unordered trees. *IEEE Transactions on Systems, Man, and Cybernetics-Part A: Systems and Humans*, *35*(2), 249-260.

[39] Romanowski, C. J., &Nagi, R. (2004). A data mining approach to forming generic bills of materials in support of variant design activities. *Journal of Computing and Information Science in Engineering*, *4*(4), 316-328.

[40] Sebzalli, Y. M., & Wang, X. Z. (2001). Knowledge discovery from process operational data using PCA and fuzzy clustering. *Engineering Applications of Artificial Intelligence*, *14*(5), 607-616.

[41] Shahbaz, M., & Harding, J. A. (2003). An Integrated data mining model for manufacturing enterprises. In *ADVANCES IN MANUFACTURING TECHNOLOGY-CONFERENCE-* (Vol. 17, pp. 539-548). TAYLOR & FRANCIS LTD.

[42] Shahbaz, M., Srinivas, M., Harding, J. A., & Turner, M. (2006). Product design and manufacturing process improvement using association rules. *Proceedings of the Institution of Mechanical Engineers, Part B: Journal of Engineering Manufacture*, *220*(2), 243-254.

[43] Shahbaz, M. (2005). *Product and Manufacturing Process Improvement Using Data Mining* (Doctoral dissertation, Loughborough University).

[44] Sim, S. K., & Chan, Y. W. (1991). A knowledge-based expert system for rolling-element bearing selection in mechanical engineering design. *Artificial Intelligence in Engineering*, *6*(3), 125-135.

[45] Song, Q. (2004). A preliminary investigation on materials informatics. *Chinese Science Bulletin*, *49*(2), 210-214.

[46] Suh, C., & Rajan, K. (2009). Invited review: data mining and informatics for crystal chemistry: establishing

IJSER

measurement techniques for mapping structure–property relationships. *Materials Science and Technology*, *25*(4), 466-471.

[47] Tan, P. N. (2006). *Introduction to data mining*. Pearson Education India.

[48] Westbrook, J. H. (2003). Materials data on the Internet. *Data Science Journal*, *2*, 198-212.

[49] Whitehall, B. L., Lu, S. Y., &Stepp, R. E. (1990). CAQ: A machine learning tool for engineering. *Artificial Intelligence in Engineering*, *5*(4), 189-198.

[50] Witten, I. H., & Frank, E. (2005). *Data Mining: Practical machine learning tools and techniques*. Morgan Kaufmann.

[51] Ye, Y., & Chiang, C. C. (2006, August). A parallel apriori algorithm for frequent itemsets mining. In *Fourth International Conference on Software Engineering Research, Management and Applications (SERA'06)* (pp. 87-94). IEEE.

IJSER

IJSER

IJSER