Rule mining in maintenance: Analysing large knowledge bases

Bernard Grabot

INP-ENIT, University of Toulouse, 47, Avenue d'Azereix, BP 1629, F-65016 Tarbes Cedex, France

ARTICLEINFO

Keywords:
Association rule mining
Maintenance
Rule visualization
Rule assessment

ABSTRACT

Association rule mining is a very powerful tool for extracting knowledge from records contained in industrial databases. A difficulty is that the mining process may result in a huge set of rules that may be difficult to analyse. This problem is often addressed by an a priori filtering of the candidate rules, that does not allow the user to have access to all the potentially interesting knowledge. Another popular solution is visual mining, where visualization techniques allow to browse through the rules. We suggest in this article a different approach generating a large number of rules as a first step, then drill-down the produced rule base using alternatively semantic analysis (based on a priori knowledge) and objective analysis (based on numerical characteristics of the rules). It will be shown on real industrial examples in the maintenance domain that UML Class Diagrams may provide an efficient support for subjective analysis, the practical management of the rules (display, sorting and filtering) being insured by a classical Spreadsheet.

1. Introduction

Companies are now computerised for a long time and have stored through years large amounts of data that could be used for generating knowledge allowing to improve their performance (Harding, Shahbaz, Srinivas, & Kusiak, 2006). The extraction of explicit knowledge from rough data is a major scientific and industrial challenge, that can be taken up by data mining techniques in a more or less automated way (Köksal, Batmaz, & Testik, 2011). Many tools can be used in that purpose, like decision trees, neural networks, genetic algorithms, k-means etc. The choice of a tool depends both on the expected type of knowledge (patterns, clusters, association rules, etc.) (Choudhary, Harding, & Tiwari, 2009) and on its potential use (clustering, diagnosis etc.) (Han and Kamber, 2006; Liao, Chu, & Hsiao, 2012). In this article, we focus on the structuration of knowledge in association rules, of the form IF (antecedents) THEN (conclusions), often considered as close to human reasoning (Koskinen, 2012). The seminal works on association rule mining have been performed on customer relationship (Agrawal, Imielinski, & Swami, 1993; Agrawal & Srikant, 1994) but these techniques can also be useful on many aspects of manufacturing systems, like quality management, production planning, production control, etc. We consider here the maintenance domain, since decreasing maintenance costs is nowadays a major industrial challenge, while the classical chain “symptoms-diagnosis-action-result” may help to structure the data bases and the extracted knowledge. As a consequence, we shall here focus on extracting knowledge from maintenance records, even if the suggested approach can be transposed to many fields.

Many articles describe rule mining algorithms, but less address the difficult problem of the analysis of the generated knowledge base, that can be very rich (Mansingh, Osei-Bryson, & Reichgelt, 2011; Potes Ruiz, Kamsu-Foguem, & Grabot, 2014). Two main approaches are used in order to cope with this problem: decreasing the number of produced rules in order to only generate “useful knowledge” (Gasmi, BenYahia, Nguifo, & Slimani, 2005) or use different ways to visualize the rules in order to explore the knowledge base. We suggest here a different approach: generate a large number of rules, without considering initially the genericity, robustness or type of generated rules, then explore the obtained (rich) rule base by performing alternatively a subjective semantic-based analysis (using already existing knowledge) and an objective analysis (based on the numerical characteristics of the rules) of the knowledge base. In order to make the method transferrable to industrial partners, we have chosen to do this using classical and simple tools, like UML and a spreadsheet. The suggested methodology is illustrated on five real case studies.

The rest of the article is organized as follows: Section 2 defines the context of industrial maintenance, summarizes the bases of rule mining and provides an overview of studies aimed at using data mining to improve maintenance. Section 3 describes our methodological proposal and the combination of tools suggested to deal with a large number of rules. These proposals are applied to actual industrial cases in Section 4, while the lessons learnt from this study are summarized in Section 5.

E-mail address: bernard.grabot@enit.fr.

https //doi.org/10.1016/j.cie.2018.11.011

2. Context and state of the art

2.1. Industrial maintenance

According to European standards (EN6:1330, 2001, 2001), maintenance is defined as “the combination of all technical, administrative and managerial actions performed during the life cycle of an item intended to retain it in, or restore it to, a state in which it can perform the required function”. A distinction is usually made between curative maintenance, aimed at restoring the resource after failure, and preventive maintenance, aimed at preventing the occurrence of defects. Preventive maintenance can be initiated at fixed intervals or conditionally. The latter is often called predictive maintenance.

Maintaining an efficient and cost-effective production system is becoming increasingly difficult as resources become more complex, particularly because of the profusion of technologies they carry (electronic, mechanical, hydraulic, pneumatic, computer, etc.) (Alsyouf, 2009). Many maintenance managers, supervisors and operators see a lack of knowledge about the plant, its equipment and processes, as the main obstacle to implementing effective maintenance procedures (Crespo Marquez and Gupta, 2006). It is therefore important to provide maintenance operators with effective decision support in their tasks of supervision, diagnosis, prognosis or in the choice of corrective or preventive actions. To do this, it is tempting to consider the masses of data accumulated over the years (Ben-Daya, Duffuaa, Raouf, Knezevic, & Ait-Kadi, 2009). Indeed, most of the large companies use for maintenance management dedicated information systems (Computer-Aided Maintenance Management Software - CMMS) or modules of their ERP (Enterprise Resource Planning) systems. These information systems store years of records on problems encountered and on the implemented solutions.

2.2. Data mining and maintenance

Data mining allows knowledge to be extracted from databases in an automatic or semi-automatic way (Köksal et al., 2011). According to Choudhary et al. (2009), however, it was initially rarely used for maintenance: in 2009, only 8% of studies using data mining in manufacturing were oriented on maintenance. The strategic importance of preventive and predictive maintenance is now acknowledged by most of the companies, and as a consequence, data mining for maintenance has experienced significant growth. Young, Fehskens, Pujara, Burger, and Edwards (2010) propose to extract rules from a database in order to link failures, causes and corrective actions to be taken in the aeronautical industry. In the same sector, Baohui, Yuxin, and Zheng-Qing (2011) seek to model the link between symptoms and corrective actions by means of rules, while Maquee, Shojaie, and Mosaddar (2012) evaluate the effectiveness of maintenance actions carried out on a bus fleet. Sammouri et al. (2012) are looking for rules combining alarms and failures. (Bin and Wensheng, 2015) uses rule mining to help diagnose high-speed trains. In (Djatna and Alitu, 2015) association rules are built to link state of resources and required maintenance actions. In (Ruschel, Alves Portela Santos, & de Freitas Rocha Loures, 2017), data mining is carried out on the time series of events generated by the process (“event logs”) to propose preventive maintenance actions. Deviations of product quality are detected in real time in Glavar, Kemeny, Nemeth, Matyas, and Monostori (2016) and linked to failures through correlations identified by data mining. In Potes Ruiz et al. (2014), data mining is used on reports of maintenance operations to generate rules linking data elements specified by the user (e.g. symptoms and cause). In Accorsi, Manzini, Pascarella, Patella, and Sassi (2017), several classification methods (K-NN, decision trees, random forest, neural networks) and rule mining (Apriori algorithm) are used to analyse the machine behaviour before failures occur.

These studies have indeed shown promising results. Despite their diversity, they nevertheless share a number of similarities. First, they seek to identify knowledge that can solve a specific problem (linking symptoms to causes, or causes to actions, or alarms to failures). Another approach could be to generate a large diversity of knowledge, then explore the obtained knowledge base for different purposes. This may possibly allow the emergence of unexpected but potentially useful knowledge. In the context of rule mining which will be ours, this raises the problem of managing a potentially very important number of rules.

In this perspective, a second resemblance between the listed studies is that, as is often the case with data mining research, they place little emphasis on the post-processing of the knowledge base (with the exception of Accorsi et al. (2017)), and in particular on knowledge analysis and validation. These two points become essential if we accept to generate large rule bases linking any type of information.

In the following section, we summarize the basics of rule mining, then show how a combination of objective and subjective analysis techniques can make useful but sometimes unforeseen knowledge emerge.

2.3. Data mining basics

An experiment can be described by a set of descriptors. A maintenance intervention can for instance be described by a date, the equipment concerned, one or several symptoms, a diagnosis and an action carried out. Recording such activity in a database D creates a “transaction”, composed of attributes (also called *types* or *data types*), each having a set of possible values. These values can be predetermined and selectable in a list (using a taxonomy of failures, symptoms or actions for instance), or freely chosen by the person describing the transaction.

Data mining aims to uncover knowledge contained in large volumes of data (Harding et al., 2006). Like statistical analysis, this knowledge may denote associations between attribute values, but also between sets of values. An *item* is the value of an attribute. An *itemset* (also called a *pattern*) is a set of items related to different attributes. A 1-itemset is an itemset containing one item; a k-itemset contains k items.

The basic principle of data mining is that knowledge can be obtained by generalizing “frequent patterns”. To measure the frequency of an itemset, its *support* can be calculated, defined as the ratio between the number of transactions in which the itemset is present and the total number of transactions contained in the database. A minimum threshold is usually defined, above which an item will be considered as *frequent*.

Different types of knowledge can be sought (Choudhary et al., 2009), among which the best known are:

- *Associations*, gathering items that often occur at the same time.
- *Patterns*, composed of item sequences staggered in time.
- *Classes*, allowing to classify transactions according to patterns known in advance.
- *Clusters*, also aiming at classification, but the groups being not known in advance.
- *Predictive knowledge*. The aim is to discover patterns leading to a prediction of the evolution of the entity observed in the transaction.

In association rule mining (Agrawal et al., 1993), knowledge is structured in association rules, “IF X THEN Y”, denoted $X \rightarrow Y$, where X and Y are frequent itemsets. X is called the antecedent and Y the consequent of the rule but it is important to notice that a rule does not express a causal link between X and Y, but only the co-occurrence of these itemsets in transactions of the database. Let us notice that if many combinations of itemsets are frequent in the data base, there is a combinatorial explosion of the number of rules that may be difficult to handle. Association rules are well adapted to the maintenance problem, since they may show a possible correlation between symptoms, failures, causes, and actions that could be useful for improving the maintenance process.

The oldest and best-known rule mining algorithm is Apriori (Agrawal and Srikant, 1994). This algorithm is executed in two steps: frequent itemsets are first identified, then rules are built by combining itemsets. The combination of itemsets during rule making is likely to provoke a combinatorial explosion of the number of rules (Gasmi, Ben Yahia, Mephu Nguifo, & Slimani, 2006). Many authors consider the generation of a large number of rules as an obstacle to the use of the knowledge base (Agrawal et al., 1993; Djatna and Alitu, 2015; Gasmi et al., 2006; Kotsiantis and Kanepoulos, 2006). In order to limit the number of rules, several solutions are possible, the basic idea being to define thresholds on numerical "measures of interest" supposed to express the relevance of a rule. The most classical measures are the *support* (already used for finding frequent items) and the *confidence*.

The support *sup* of an association rule X → Y is the proportion of transactions in D that contains both X and Y (Eq. (1)).

$$\text{sup} = P(X\cap Y) = (\text{number of transactions containing both X and Y}) / (\text{total number of transactions}) \qquad (1)$$

The confidence *conf* of the association rule X → Y is a measure of the reliability of the rule, determined by the percentage of transactions in D containing X that also contain Y (Larose, 2005). Confidence is the conditional probability P (Y|X) (Eq. (2)).

$$\text{conf} \quad P(Y|X) = P(X\cap Y)/P(X) = (\text{number of transactions containing both X and Y}) / (\text{number of transactions containing X}) \qquad (2)$$

Minimum values of the support (*minsup*) and confidence (*minconf*) can be defined: the rules that do not reach these thresholds are removed by the mining algorithm (Agrawal et al., 1993). An a priori choice of thresholds may seem arbitrary: some authors consider that it is more natural to choose the number of desired rules (Garcia, Romero, Ventura, & de Castro, 2009), once the rules have been produced and ranked according to the chosen measure(s) (Scheffer, 2005). Another strategy is to eliminate the "long rules", containing many items in the IF and THEN parts (Palanisamy, 2006; Scheffer, 2005).

Many other measures have been defined since the initial work of Agrawal et al., either for measuring the generality of a correlation (Laplace, chi-square statistics, correlation coefficients, entropy gain, interest, conviction, etc. (Garcia et al., 2009)) or on the opposite its specificity (peculiarity, diversity, novelty, surprisingness… (Geng and Hamilton, 2006). Indeed, interesting knowledge may also come from rare associations, often called "outliers".

More generally, for Geng and Hamilton (2006), several types of measures can be distinguished to assess the value of a knowledge base:

- objective measures, only based on the raw data, like support and confidence,
- subjective measures, taking into account both the data and the user's expectations. Potes Ruiz et al. (2014) suggest for instance to only keep the rules consistent with a pattern defined using conceptual graphs.
- semantic-based measures, involving domain knowledge provided by the user. Because semantic measures involve domain knowledge provided by the user, some researchers consider them a special type of subjective measures (Yao, Chen, & Yang, 2006).

On a medical application, Mansingh et al. (2011) control for instance the extraction of the rules using both a domain ontology (allowing a subjective analysis) and an objective measure expressing the strength of a rule.

A semantic-based analysis of a rule may be facilitated by considering the strength of the link between antecedent and consequent of the rule. In that purpose, another classical measure, the *lift*, may be of interest. The lift (or *interest factor*) can be obtained by dividing the

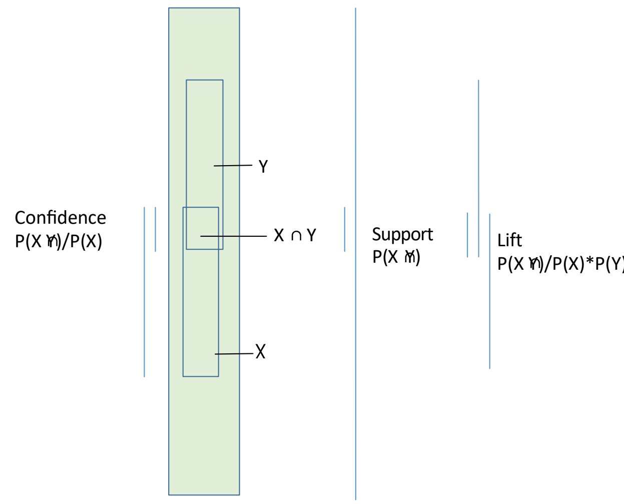

Fig. 1. Support, confidence and lift.

confidence of a rule by the unconditional probability of the consequent, or by dividing the support by the probability of the antecedent times the probability of the consequent:

$$\text{lift}(X \rightarrow Y) \quad P(Y/X)/P(Y) \quad \text{support}/(P(X).P(Y)) \qquad (3)$$

The interpretation of the lift is as follows:

- if lift = 1, X and Y are independent,
- if lift > 1: X and Y are positively correlated,
- if lift < 1: X and Y are negatively correlated.

Fig. 1 summarizes the way the support, confidence and lift are calculated, emphasising their complementarity. The support does not take into account the frequencies of X and Y (only their intersection; middle part of Fig. 1) while the confidence does not take into account the frequency of Y (left part of Fig. 1). Tan, Steinbach, and Kumar (2006) consider for instance that high confidence rules can be misleading, because the confidence measure ignores the support of the itemset appearing in the rule consequent. This is done in the lift (right part of the figure). Although classical, these three measures allow to give a rather comprehensive view on a produced rule within its context. Since our objective is more on the methodology than on the measures themselves, we have adopted these classical measures for our study.

Very often, (with the notable exception of Mansingh et al. (2011)), objective evaluation is performed first, since it can easily be embedded in the mining algorithms, whereas semantic and subjective evaluation are performed on the remaining set of rules. As a consequence, rules that do not meet the thresholds during the objective evaluation phase are usually not considered by the experts of the field. On the opposite, we suggest to accept the possibility that many rules are created for accessing a wide variety of knowledge. We shall see in next section how a large set of rules can be managed through visualization techniques.

2.4. *Visualizing association rules*

Based on the observation that the analysis of many association rules may be difficult (Hahsler and Chelluboina, 2015), some authors suggest to address the analysis of large sets of rules through visualization techniques (see surveys in Bruzzese and Davino (2008) or Hahsler et al. (2015)). Without claiming to be exhaustive (especially since these representations have been object of many variants), we can list:

- *two-dimensional matrices*: the antecedent items are positioned on one axis and the consequent items on the other. The height and colour of

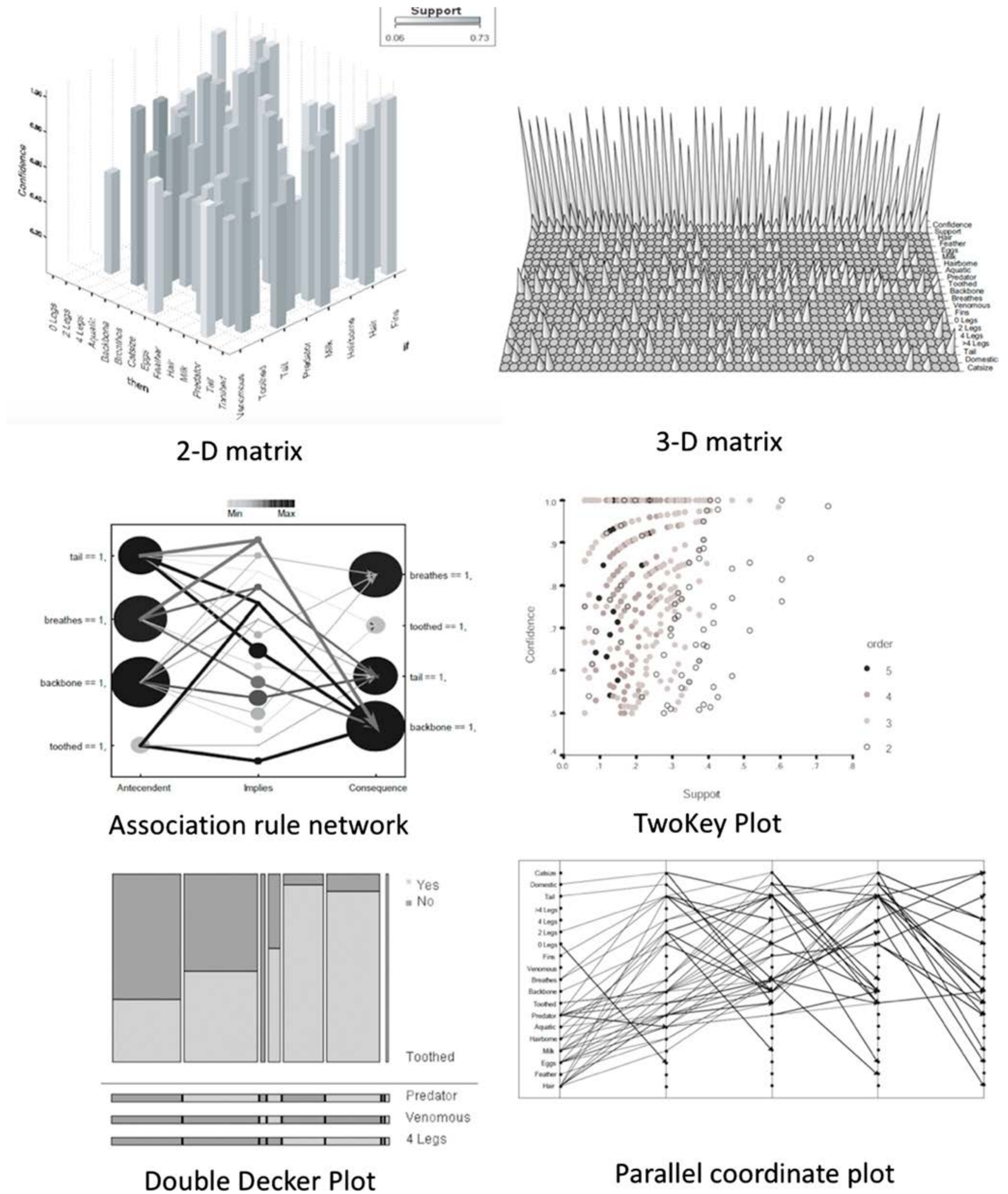

Fig. 2. Visualization of association rules (Bruzzese & Davino, 2008).

a bar represent the support and confidence. These techniques hardly allow to represent many-to-one relationships.

– *3D-visualization*: the rows of a matrix represent the items while the columns represent the rules. Bars with different heights visualize the antecedent and consequent parts while other bars represent the support and confidence of each rule.

– *Association rule networks*: each node represents an item, and the edges represent associations between items. The support and confidence are represented by the colour and width of the arrows.

– *TwoKey plot*: the x-axis and y-axis represent respectively the support and confidence; a rule is represented by a point on the graph.

– *Mosaic plots* and *Double decker plots*, representing a rule and its

related rules by bar graphs.
– in *Parallel coordinate plots*, items are in lines, a rule being represented by joining antecedents by arrows and consequents by lines.

Examples of these visualizations, provided in Bruzzese and Davino (2008), are shown in Fig. 2.

Let us notice that the visualization methods based on the items may be difficult to handle when many items are present in the rules, in relation with the combinatorial explosion already discussed. When the number of rules is important, a factorial method may help to synthetize the information stored in the rules (Bruzzese & Davino, 2008) with some loss of information. The result may be represented on factorial planes for items, rules, or joint representation.

These methods may indeed allow to represent large number of rules (several hundreds to several thousand) but the TwoKey plot is the easier tool for taking into account hundreds of thousands of rules, which is the order of magnitude that we shall deal with.

3. Suggested method

A preliminary proposal was made in Grabot (2017): (i) extracting as much knowledge as possible from maintenance reports, (ii) display the rules obtained in a spreadsheet and (iii) explore the obtained rule base using the spreadsheet filters, by formulating hypotheses on potentially interesting knowledge, with the support of a class model of the database. This method proved to be promising on the basis of initial experiments, allowing in particular to efficiently manage hundreds of thousands of rules using a standard spreadsheet. However, more extensive tests have shown that very different types of knowledge bases can be obtained. An objective analysis based on a graphical representation of the rules can effectively complement this first semantic analysis.

3.1. Knowledge base generation

Numerous open source or commercial software programs allow the use of data mining tools on an existing database, including the R language that offers a very complete environment of statistical and graphical processing.[1] For our part, we have chosen to encapsulate data mining algorithms provided by Philippe Fournier-Viger[2] in in-house developments for reasons of competence and flexibility, but this choice does not condition the use of the method.

Our objective is to facilitate the post-processing of the resulting rule base, not to propose a particular algorithm for extracting the rules. We have therefore chosen the most classic algorithm, Apriori (Agrawal et al., 1993), for the extraction of rules, as well as classical measures (support, confidence and lift) for the objective evaluation of the rules. Creating the rule base requires several steps:

Creation of the database: In order to have files that are easy to manipulate in any environment, we have worked on csv files coming from Excel exports from the databases of the companies that provided the data for tests. All the data bases analysed in Section 4 are extractions from the SAP ECC Maintenance module, which is understandable since SAP ECC is the ERP most often encountered in large companies.

Data cleaning: Preparing the data for a data mining process is recognised as a difficult and long-lasting task. A structured process for data-cleaning is for instance suggested in Mansingh, Osei-Bryson, Rao, and McNaughton (2016). As a first step, some attributes can be removed without loss of information: for instance, the reference number of the manufacturing order aiming at identifying it in a unique way can obviously not be object of any generalisation. Other attributes that make a transaction too specific should also be pre-processed: for instance, several databases mention several dates for each maintenance activity (planned, realized, etc.). Such information can hardly be generalized, since few transactions concern the same day. As a consequence, the dates have been grouped by periods for our experiments. In that purpose, a simple program was developed allowing to group dates by week, fortnight or month, depending on the number of transactions per period.

As a third step, specific attention must be paid to text fields filled by the actors in a free manner, like the description of failures or causes for some companies. Indeed, many expressions can be used to describe the same observations (not mentioning spelling errors). Such fields should obviously be homogenized to allow generalisation. This is in practice only possible in a manual way, and so on a limited number of transactions, if advanced text mining algorithms are not used (see (Arif-Uz-Saman, Cholette, Ma, & Karim, 2016) for an example of the use of text mining on the maintenance domain). During this study, some fields, like the symptoms for instance, have been homogenized using taxonomies, but it has only been done on the small databases. On the largest databases, the text fields have been removed in order to simplify the first tests, described hereafter.

Extraction of the rules base: As stated in Section 2, the Apriori algorithm first looks for frequent itemsets in relation to a minimum support. In order to generate as many rules as possible, we propose to set the three thresholds on support, confidence and lift measurements at values low enough to generate the highest number of rules that can be displayed and manipulated by a spreadsheet. In practice, this number depends not only on the spreadsheet but also on the RAM of the computer used. With a conventional 8 GB memory computer, the limit of the number of rules that can be displayed is below the million for Excel, and is even higher with OpenOffice. Filters are then applied to the different attributes (columns in the spreadsheet).

3.2. Result analysis

As shown in Section 2, most studies published on rule mining in maintenance have a specific objective: to link symptoms to causes or causes to actions for example. Combined with the choice of fairly high thresholds for the measures of interest, the selection of the corresponding attributes makes it possible to drastically limit the number of rules produced, so the number of rules to assess. In our case, it is necessary to provide a framework for the subsequent analysis of the generated rules. We suggest to combine objective analysis (related to the measures of interest) and subjective analysis (related to the analyst's interest and to prior knowledge) in that purpose. To do this, we propose the approach summarized in Fig. 3 using the SADT syntax, based on the idea of using only simple tools.

Ontologies aiming at describing the whole domain of maintenance have already been suggested (see for instance (Karray, Chebel-Morello, & Zerhouni, 2011)). We only need here a simple reference model of the record of a maintenance order. We have therefore proposed in Grabot (2017) to use a UML class model (Fowler, 2004) for providing such simple reference model in order to facilitate subjective analysis (activity (S) in Fig. 3). Fig. 4 shows this basic model. This class diagram can be used as a “map” for choosing possible “paths” between data using the relationships between attributes and their cardinalities. For instance, the diagram of Fig. 4 suggests that there could exist associations between symptoms, failures and causes, or between maintenance orders, failures and equipments. Considering the cardinalities is of great interest in that purpose. For instance, the cardinality “*/*” between failure and symptom suggests (i) that the same symptom can be associated to different failures, (ii) that the same failure can result in different symptoms. It is easy to check this assumption using the filters on the itemsets of the rule: it is enough to select a failure in the antecedent part of a rule, and to check whether several symptoms appear in the filter of the consequent part of the rule. This helps to check whether the

[1] https //www.r-project.org/.
[2] http //www.philippe-fournier-viger.com/spmf/index.php.

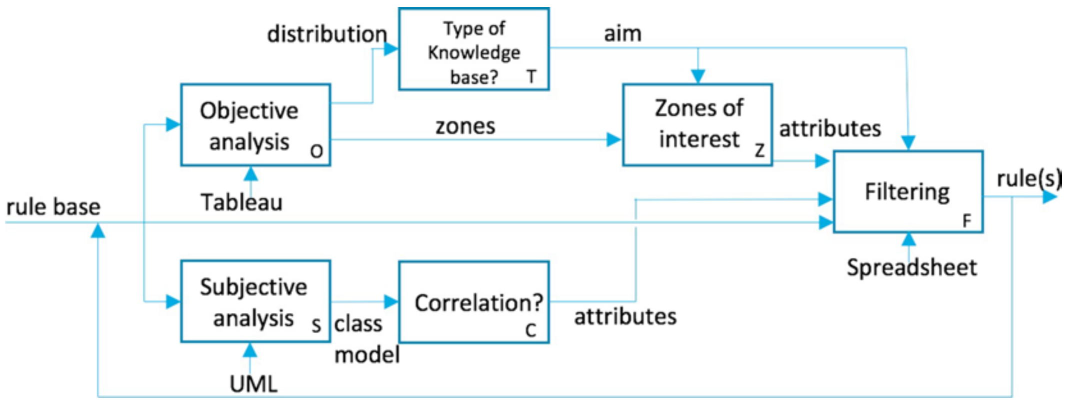

Fig. 3. Suggested methodology.

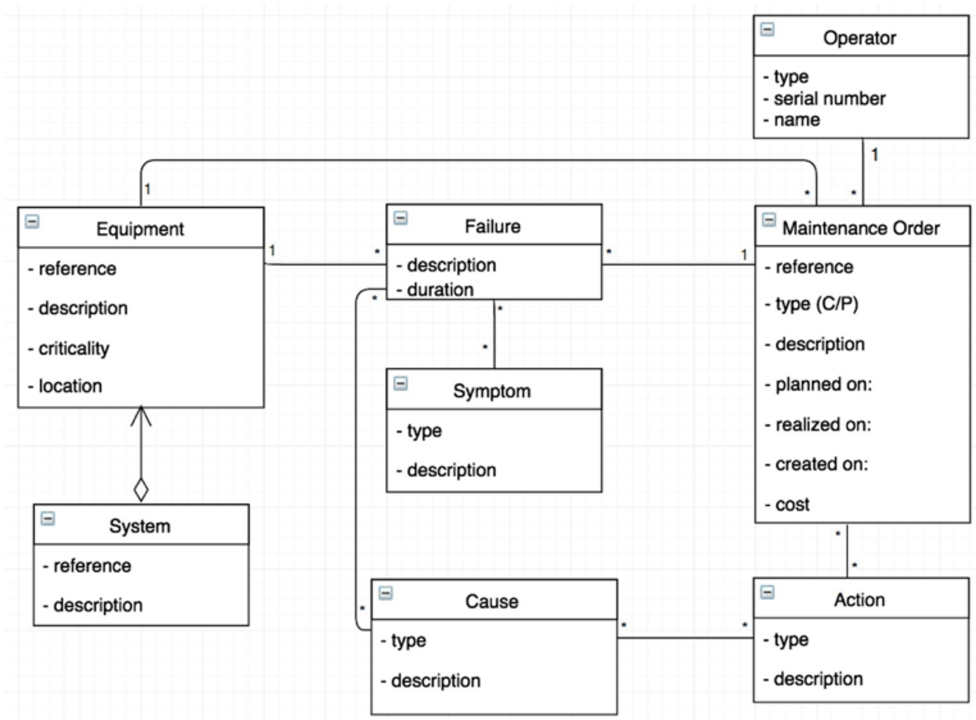

Fig. 4. Basic model of a maintenance order (Grabot, 2017).

theoretical view on the system is consistent with the operational reality of the process.

Since the databases that we will consider in the case studies all come from the same ERP (SAP ECC), one might expect a single reference model to be sufficient. This is not the case: the first reason is that SAP ECC has pre-set versions for a large number of industrial areas. The maintenance module of the “Aerospace and Defence” version, dedicated to discrete manufacturing, is for instance very different from the one of the “Pharmaceuticals” version, oriented towards continuous processes. The second reason is that the stored attributes and the database extractions can be customized to specific requirements, which has been done by all the analysed companies. Starting from a standard class model (Fig. 4) positioning the classical concepts in maintenance management (mainly: the maintenance order describing the equipment, symptoms, the diagnosed cause, and the action carried out), it will thus be required to produce the real model used in each company, which can be significantly different. This model will make it possible to formulate hypotheses of correlation between attributes ((C) in Fig. 3) allowing to define sequences of use of the spreadsheet filters. As illustrated in Section 4, trying to find associations between two concepts using the class model of Fig. 4 is rather easy: it is enough to consider two concepts linked by a relation and to look in the spreadsheet whether values of the two items can be respectively found in the IF and THEN part of some rules. Extending the reasoning to more than two concepts is a bit more difficult because the items can be split in different manners in the two parts of the rules, but it can nevertheless be done without major problem, even if, for better clarity, the examples of Section 4 focus on two items.

Our initial tests have shown that this “local” approach, guided by the analyst's prior knowledge, rarely leads to unexpected knowledge, that may be the most interesting. We therefore suggest that this subjective analysis could be flexibly combined with a more comprehensive objective analysis based on measures of interest. This should make it possible to identify groups of potentially interesting rules according to what the user is looking for. We propose for that a two-stage approach that will be detailed on the case studies in Section 4:

- The knowledge base is firstly mapped by means of two main graphical tools visualizing the distribution density of the rules according to the different measures of interest, and the links between

Table 1
Main characteristics of the companies and of their databases.

	Domain	Type of maintenance	Number of transactions	Period	Number of attributes before cleaning	Number of attributes after cleaning
Aero	Aeronautics (assembly-painting)	Maintenance on discrete process; curative	262	3 months	23	13
Auto	Automotive (electronical sub-systems)	Maintenance on discrete process; curative and preventive	4312	1 year	14	10
Pharm	Pharmaceutics	Maintenance on continuous processes; curative	537	1 year	20	14
Maint	Maintenance for aeronautics	Maintenance on discrete process; traceability for contractual constraints	13,000	7 years	43	23
Sem	Semiconductors	Maintenance on discrete process; curative and preventive	14,524	2 years	20	11

measures of interest (confidence as a function of support and lift as a function of trust) ((C) in Fig. 3).

- After analysing the characteristics of the knowledge base ((T) in Fig. 3), it is possible to identify areas of interest in the knowledge base (Z), the choice of which defines the objectives of the analysis (looking for generic knowledge or rare knowledge, for example).

The attribute selection, which is the result of the objective and subjective analyses, will allow to manage the spreadsheet filters for selecting potentially interesting rules (F).

As it will be shown in Section 4, it may be useful to perform objective and subjective analysis alternatively in order to explore the data base (denoted by a feedback loop between activity (F) and activities (O) and (S) in Fig. 4).

4. Case studies

Five companies (of which the names have been changed) have accepted to provide real data for testing the method. Table 1 summarizes the sector in which the companies operate and the main characteristics of the provided data bases. Let us notice that the provided files are extractions of the main database, and only contain the attributes chosen for the tests by the company.

Section 4.1 gives some details on the data cleaning of the data bases used for the tests. As explained in Section 2, the rules are then generated using a standard Apriori algorithm, used with low thresholds on the measures of interest (Section 4.2). For analysing the obtained rule base, the user may look for rules having given characteristics (generic rules or exceptions; robust rules etc.) or may perform some hypothesis on correlations between attributes, then try to validate them. In the first case, the analysis should begin by an objective analysis ((O) in Fig. 3; see Sections 4.4 and 4.5), in the second case by a subjective analysis ((S) in Fig. 3; see Section 4.6).

4.1. Data cleaning

Only a basic data cleaning process was performed. Basically, the text fields were replaced by discrete sets of values using taxonomies suggested by the maintenance experts in the small data bases. For instances, “free” sentences describing symptoms were replaced by a type of symptom taken from a standard list. Since performing manually this operation is highly time consuming, it was decided to conduct the first tests on the large databases after removing all the text fields.

Since our algorithm replaces an empty content by the content “EMPTY”, empty fields were processed as a specific value of the considered attributes. As shown in Section 4, the result is that empty fields are involved in many association rules, creating unexpected knowledge on the way the information was filled.

4.2. UML class diagrams

The first step is to build the UML class diagrams that will be used for activity (S) of Fig. 3. Fig. 5 shows the class diagrams of the maintenance data base in the five enterprises. For Aero, a rather simple model can be noticed, directly connecting a work order to a symptom and an equipment. The failure is not distinguished from its symptoms. The attributes removed during the cleaning phase essentially relate to planned and real action dates.

For Auto, the model is centred on a failure, in which a maintenance action is assimilated to a part change. Again, failure and symptom are not formally distinguished.

For Pharm, Maint and Sem, more standard models distinguish failure (called damage by Pharm), cause, action and maintenance order. It should be noticed that Maint being a company providing a maintenance service, many attributes are intended to justify that the conditions of the contract are met. For these tests, we simplified the data

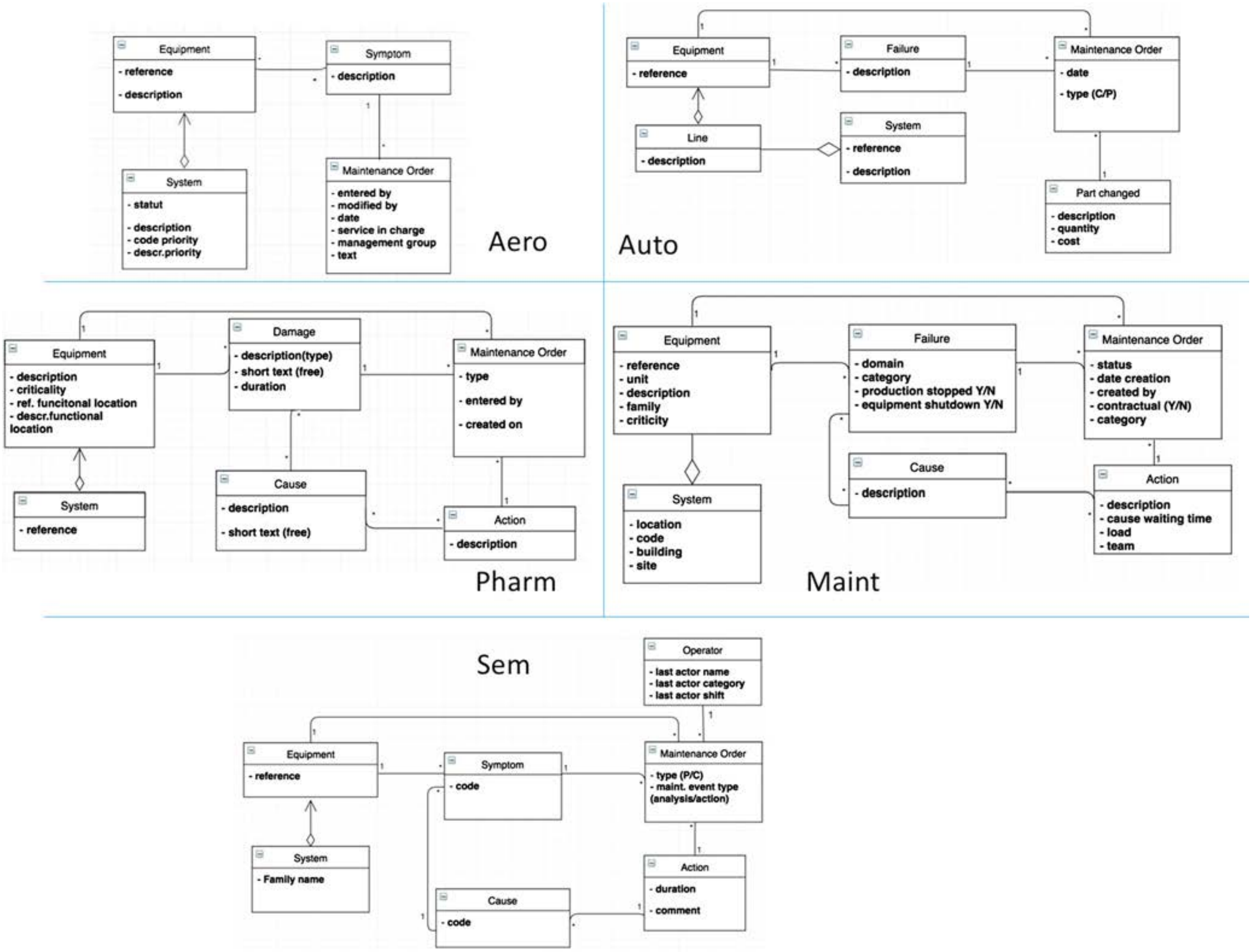

Fig. 5. Class diagrams.

base by removing a significant number of dates (date of failure, intervention data, date of temporary repair, date of return to a nominal state, etc.), which could also be analysed.

4.3. Rules extraction

Several tests have been conducted in order to extract a rich and multipurpose knowledge base. For that, the minsup and minconf have been set to very low values (for generating both generic and rare rules), while the minlift has been set to 0, in order to allow the generation of rules involving attributes of any types of correlation. The choice of the thresholds, and as a consequence the number of rules, was limited either by the memory used by the rule mining software, or by the possibilities of the spreadsheets used to display the rules and to act on them through scrolls and filters. The set-ups of the mining experiments so that the obtained rule base are detailed in Table 2, were "Items" is the number of different values of attributes found in the data base. Let us note for instance that for the set up (minsup = 1, minconf = 1, minlift = 0) for the Aero database, 2 601 506 rules are generated by the data mining software and stored in a text file without problem, but Excel cannot open the file, while LibreOffice "only" opens the first 1 048 576 rules. In any case, the problem is not only to open the file but also to manipulate the rules, which is why we have in each case generated less than 1 million rules.

Table 2
Mining the rules.

	Attributes	Transactions	Items	Frequent itemsets	Minsup, minconf, minlift	Number of rules
Aero	13	262	446	35,997	5 5 0	131548
Auto	10	4312	5284	1354	1 1 0	12260
Pharm	14	537	2148	4668	1 1 0	103648
Maint	23	13,000	12,724	10,656	8 8 0	645348
ST	11	14,524	2447	4663	1 1 0	66582

Examples of rules displayed by the spreadsheet are shown in Fig. 6 (here rules with only one antecedent and one consequent). There is one rule per line in Fig. 6, the field "Cond1" denoting the IF part of the rule and "Resul1" the THEN part. "nbr" denotes the number of transactions on which the rule is based (the support also provides this information as a ratio) while the support, confidence and lift are the one of each rule.

4.4. Characterisation of the rule bases

Let us consider steps (O) and (T) of Fig. 3. A first point is to determine the objective of the analysis, i.e. whether we shall mainly look for generic or specific rules (outliers). In that purpose, it is interesting to first consider the distribution of the rules according to their measures of interest (see Fig. 7). Let us remind that the support gives the percentage of transactions expressed by a rule, the confidence whether a rule is

	A	B	H	I	O	P	Q	R
1	rule nbr	Cond 1	symbol	Resul 1	nbr	support	confidenc	lift
2	1	FAMILY_NAME=DGF-USG	==>	MAINT_TYPE=CM	13887	0,96	0,96	1,00
3	2	MAINT_TYPE=CM	==>	FAMILY_NAME=DGF-USG	13887	0,96	1,00	1,00
4	3	EQUIPMENT_NAME=DGF10C	==>	MAINT_TYPE=CM	831	0,06	0,96	1,00
5	4	MAINT_TYPE=CM	==>	EQUIPMENT_NAME=DGF10C	831	0,06	0,06	1,00

Fig. 6. Examples of rules.

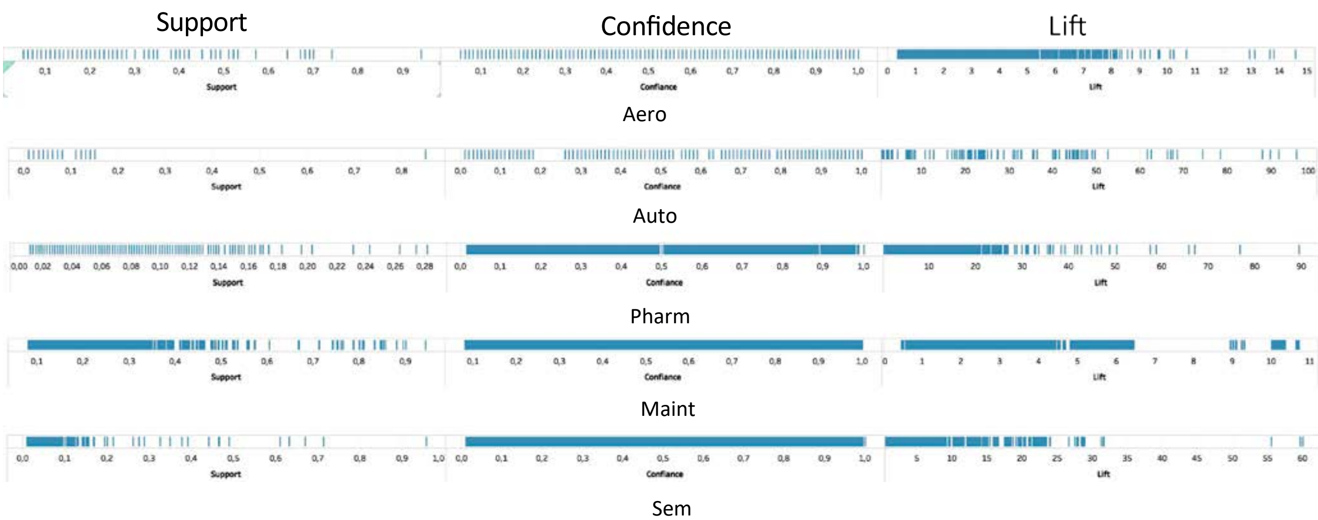

Fig. 7. Characterisation of the rule bases.

"robust" or not (the antecedent and consequent are often present at the same time), while the lift allows to analyse more precisely the link between antecedent and consequent (negative correlation; independence or positive correlation).

Let us remind that lift = P(X ∩ Y)/P(X)*P(Y). If X and Y always appear at the same time, P(X)=P(Y)=P(X ∩ Y), so lift = 1/P(X)=1/P(Y). The lift is therefore maximum when X and Y always occur at the same time, and are rare. This means that high values of the lift denote rare events: the so-called "outliers" that can easily go unnoticed by traditional techniques filtering the rules according to the support. This can be easily checked on the rule bases: if only the rules of support > 0.3 are selected, all the displayed rules have lifts and confidence around 1, while all the rules with high lift have a support inferior to 0.3.

It can for instance be seen in Fig. 7 that all the rules bases have a relatively well distributed confidence (i.e. rules of any support can be found in the rule base, i.e. rules with all possible degrees of genericity/specificity). Nevertheless, Fig. 7 allows to define what is promising for each rule base:

- Auto is mainly composed of rules of low support (rare knowledge) but the distribution of confidence is regular, meaning that robust knowledge should be present. Moreover, the maximum lift being rather high (100) with a pretty regular distribution, it is likely that high correlations are present in these rare pieces of knowledge.
- Aero and Maint have close characteristics with rather well distributed supports (denoting rules of various levels of genericity/specificity), well distributed confidences (i.e. from robust to exceptional rules) and relatively low maximum correlations (resp. 15 and 11) showing that the positive correlations do not concern outliers. It is also interesting to notice that few highly negative correlations are absent (the minimum lift is 0.35 for Aero and 0.46 for Maint).
- Sem also includes mainly rules with low supports, some of them having a rather important lift (60), denoting some rare events with high correlations that deserve to be analysed.
- Pharm is only composed of rules of low support (the maximum support is 0.28 in Fig. 7 whereas it is around 0.9 for all the other bases) with a maximum lift at 90 and both positively and negatively correlated items. This is again a promising field for finding rare and robust knowledge, but not very generic knowledge (coming from rules of high support, i.e. knowledge generalized from many transactions).

4.5. *Rules and groups of rules of specific interest*

It is now interesting to go deeper in the analysis, and to identify groups of promising rules. Fig. 8 positions the rules with their support on the x-axis and their confidence on the y-axis, the size of the point denoting the lift. Such representation can be obtained using a classical spreadsheet but the use of a tool dedicated to data visualisation, Tableau[3], has easily allowed to represent the lift of a rule by the size of the point, and to display a rule "*in extenso*" by positioning the mouse on the point that represent it (see the graph of Pharm in Fig. 8).

A first comment when looking at the curves of Fig. 8 is that all the points are above the line support = confidence. Indeed, confidence > support since confidence = support/P(X) with 0 < P(X) < 1.

The points denoting the rules are initially distributed on straight lines (bottom points). This is also easy to explain: the x-axis (support) represents the number of transactions concerned. The increment on the x-axis is therefore 1/n, n being the total number of transactions of the data base. For instance, this increment is 1/537 = 0.00186 for Pharm as it can be verified on the curve. When browsing points on the same

[3] https //www.tableau.com/.

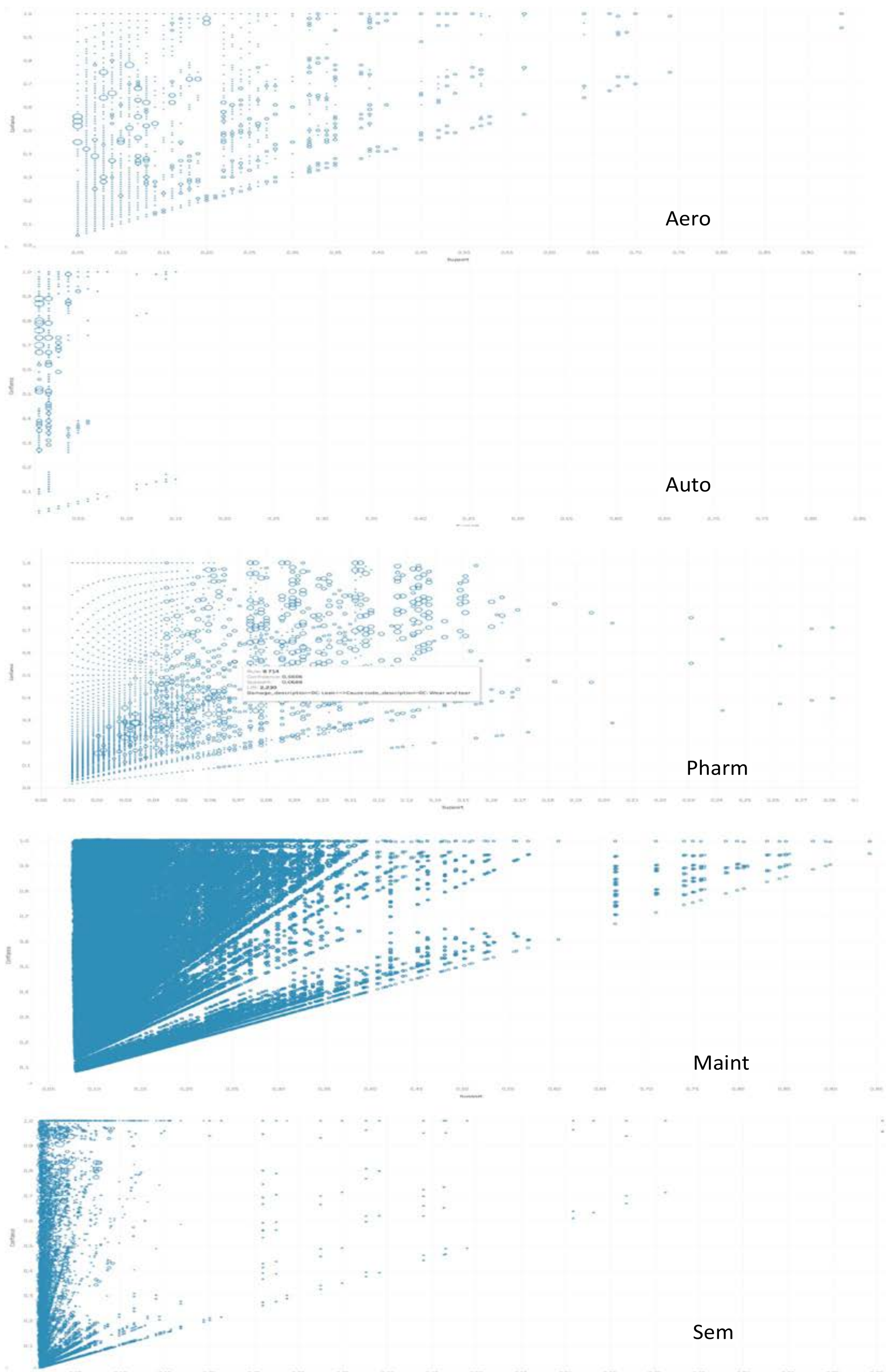

Fig. 8. Confidence = f(Support).

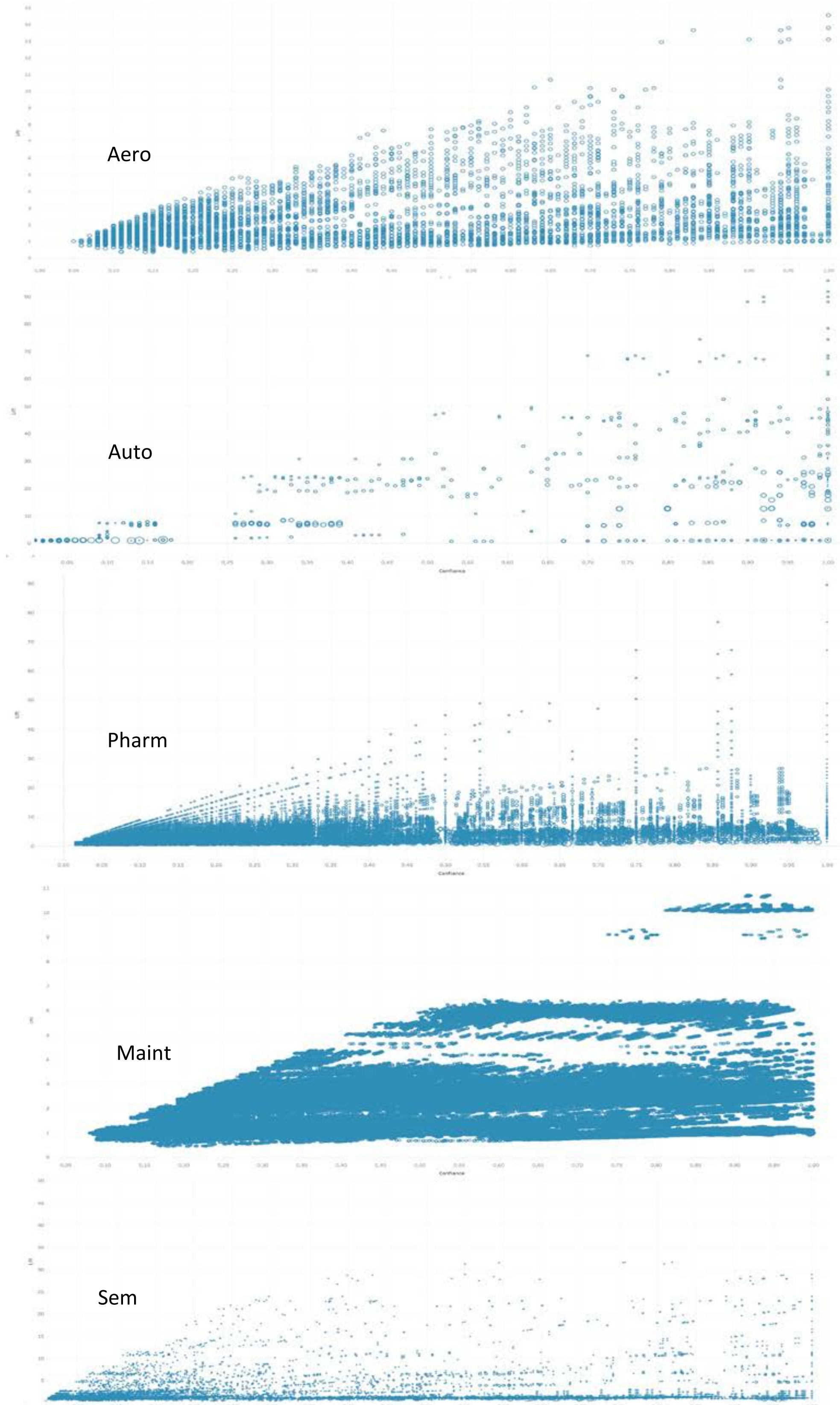

Fig. 9. Lift = f(confidence).

line, it can be seen that the corresponding rules have the same antecedent X, so P(X) is a constant. Since confidence = support/P(X), 1/P(X) is the slope of the line.

When the confidence is higher, the lines turn into regular curves (see Pharm in Fig. 8). The explanation is that between two points, the support $P(X \cap Y)$ increases of one transaction, i.e. $(1/n) = a$; P(X) also increases of $(1/n) = a$. Let us write conf $= P(X \cap Y)/P(X) = x/y$ with $x < y$ for simplification. The confidence of the following point is $(x + a)/(y + a)$, then $(x + 2a)/(y + 2a)$ for the next point, etc.

It is easy to demonstrate that the increment of confidence decreases at each (regular) increment of support, and has for limit 0, which explains the curves (same increment on the x-axis, decreasing increment on the y-axis).

Fig. 9 shows the rule bases visualised with the confidence on the x-axis and the lift on the y-axis, the support being represented by the size of the points.

The lift can also be obtained by dividing the confidence of a rule by the unconditional probability of the consequent, therefore lift > conf, that can be verified on Fig. 9.

The rules at the top of each graphic are rules of high lift, and are much easier to detect here than on the graphics of Fig. 8. It is again possible to see that Sem and Pharm contain rules with very low lift (around 0), denoting interesting negative correlations that will be analysed. Three groups of rules appear clearly on the top of the figure related to Maint, that also deserve to be investigated.

4.6. *Examples of findings*

Combining semantic analysis and objective analysis for exploring an entire rule base takes time and space for giving exhaustive explanations. We shall only show here some illustrative examples of quite different ways to combine these tools.

Aero:

Objective analysis: A first natural idea when looking at the Aero graph of Fig. 8 is to analyse the two rules with high support and confidence (top-right corner of the graph). They are two symmetrical rules:

Service_in_charge="MAINTTLP"= > Entered_by="N425201″ sup = 0.94 conf = 0.94 lift = 1

Entered_by="N425201″= > Service_in_charge="MAINTTLP" sup = 0.94 conf = 1 lift = 1

These rules mean:

- that the worker N425201 belongs to the service MAINTTLP (lift = 1)
- that almost all the transactions are entered by this person (sup = 0.94)
- that all the transactions entered by N425201 concern service MAINTTLP (conf = 1 for the second rule) but that in some rare cases, somebody else than N425201 has entered an order for service MAINTTLP (conf = 0.94 for service MAINTTLP).

These findings allowed to detect an anomaly, N425201 being the maintenance actor in charge of the service. Switching to a subjective analysis using the filters of the spreadsheet, it is immediately possible to see that the other person having entered orders is N400913.

By checking the possible values of the antecedent, it can be quickly seen that only N400913 has modified orders, which is a valuable information denoting another anomaly.

More generally, it can be checked by browsing among the rules of high confidence and support that most of these rules deal with the fact that the orders have mainly be created by N425201, and always modified by N400913.

Looking for outliers can either be performed by browsing the rules of high lift (large bubbles in Fig. 8 or points at the top right corner in Fig. 9). The information displayed shows that several rules correspond to each point (denoted by * for multiple instances in the rule number and attributes). Moving to the spreadsheet is therefore preferable, since it allows to see that many rules (8658 on 131548) have a maximum lift (here 14.56), all related to the same 18 transactions. These rules describe the same failure occurring on a paint car, mentioned 18 times in the database on the same day: this outlier clearly denotes a database entry problem, that would not be spontaneously investigated by the analyst.

Auto:

Objective analysis: It is again easy to check that, with almost only rules of very low support, Auto can be used to find outliers but not generic knowledge.

Subjective analysis: The class diagram is more interesting, and suggests for instance to investigate a possible link between part changed, failure and equipment.

A great number of components are mentioned in the transactions but only five are present in the rules, meaning that they concern at least 1% of the transactions. The most often mentioned component is involved in 64 transactions, but no rule mentions the equipment on which the component is mounted, meaning that the support threshold of 1% is not reached: the article can be mounted on many devices.

As shown by Figs. 8 and 9, the structure of the rule base is atypical and denotes highly scattered transactions: only two symmetrical rules have supports higher than 0.15. They express that, for fixing a failure, the quantity of replacement parts is "1" in 85% of the transactions. Such finding has a poor interest by itself, but becomes more informative when the two symmetrical rules are considered: the rule Order_type="curative"= > quant. = "1" has a confidence of 0.99 while the reverse rule has a confidence of 0.86, meaning that preventive operations require several components more often than curative ones. This unexpected result also deserves analysis since it may suggest that less components could be changed during the preventive maintenance activities.

Pharm:

Objective analysis: The rules of high support in Fig. 8 mainly include current values of attributes like equipment criticality and order type, which is not very informative. A couple of symmetrical rules with a support of 0.23 are more interesting:

Damage_description=0C: Leak= > Object_part_Descript = 0C: Piping and fittings

with sup = 0.23 conf = 0.76 and lift = 1.91

and:

Object_part_Descript=0C: Piping and fittings= > Damage_description=0C: Leak

with sup = 0.23 conf = 0.55 and lift = 1.81

These rules show that leaks of piping and fittings concern 23% of the transactions denoting a recurrent problem; 76% of the leaks come from piping and fittings (first rule) but piping and fitting have other problems in 21% of the transactions that concern them (0.76 minus 0.55). By exploring the rule base, it can be easily found that this corresponds to "mechanical damages" that do not result in leaks. This shows again the interest to interpret the symmetrical rules as a whole and not separately.

It can be noticed that most of the rules located in the central part of Fig. 8 have "empty" as the value of some of their attributes, illustrating that a common point to many transactions is that the forms are not completed correctly. By switching to the spreadsheet, it is possible to select an operator, then check whether many rules have been generated joining his name and empty fields. One of the operators let for instance the field "cause" empty three times more often than the other operators. This may denote a problem of this operator in the diagnostic phase that should be fixed through training.

Subjective analysis: This database provides much more information, as denoted by the class diagram of Fig. 5. The class diagram suggests that it can be of interest to study the possible links between the related concepts equipment-damage-cause-action, according to their descriptions entered using a taxonomy (the short texts would require text analysis).

Fig. 10. Superimposition of rules.

The comparison between the breakdown duration and the criticality of the equipment is also of interest, showing that very few critical machines need more than 3 h to be repaired but that the number of machines of “high” criticality having been repaired in 2 h is the same than for machines of “medium” criticality.

It can also be noticed that one rule links the cause description “operator error” and the criticality of an equipment. This rule concerns an equipment of “high” criticality, denoting that errors occur more often on highly critical machines. This paradoxical result suggests to improve the training of the operators on highly critical equipment, often more complex than the others.

Maint:

Objective analysis: A lot of rules with support = 0.6662 can be distinguished on the graph of Fig. 8. When using the spreadsheet to select them, it can be seen that 30 rules link no production shutdown, curative maintenance, under contract and status = closed. This means that many curative maintenance operations have been conducted without stopping production, which may set into question the interest of preventive maintenance on some specific machines.

The two rules on the top right corner of Fig. 8 state that the waiting cause is empty for most contractual failures. Then, many rules include the “closed” status, shared by most transactions.

The graph of Fig. 9 is more interesting, since it allows to distinguish three groups of rules on the top of the graph. Nevertheless, these groups are difficult to explore using the graphs since, even when zooming, many rules are superimposed and cannot be visualized (see the “*” in Fig. 10 denoting multiple different attributes or values). This shows the limits of the visualization of such a number of rules on the same graphic.

Subjective analysis: The Maint class model is one of the most comprehensive, offering many opportunities to define a path of interest between the attributes of the objects. On the base of the model of Fig. 5, it can for instance be interesting to try to link:

- waiting causes to causes of failures or failures,
- production shutdown to equipment or failures,
- equipment to failure domains, etc.

There are 15 different waiting causes in the data base, but none of them reaches the 8% support threshold that would allow to create a rule.

Indeed, Waiting_cause=”empty” is the only value present in the rules (representing 12,310 transactions on 13000). This shows the interest to be able to still decrease the minsup, and to manage more rules.

Failures have resulted in production stoppages in 1240 transactions. 15 rules thus include a production stoppage, among which the most interesting ones are the two symmetrical rules:

Production_stop=”+”= > Non-contractual=”-“ sup = 0.1 conf = 1 lift = 1

Non-contractual=”-“= > Production_stop=”+” sup = 0.1 conf = 0.1 lift = 1

The first indicates that no production stoppage is a “non-contractual” breakdown, which is a good indicator for the company for which Maint works: indeed, the critical resources are covered by the maintenance contract (let us remind that the company provides maintenance services at a customer’s site).

The second shows that only 10% of contractual breakdowns result in a shutdown. The selection of a “vital” level of criticity as antecedent of the rules allows to obtain 957 rules describing failures on vital equipment. The consequent parts of these rules allow to check that none of these failures resulted in stopping the machine, and none was non-contractual, which is a very good result that can be used by Maint to demonstrate its effectiveness to its customer.

24% of the failures on vital equipment concern the same domain (electrical, electromechanical, electronic). Improving the availability on vital equipment clearly requires an investigation on this sector.

The rules with confidence = 1 (11032) mainly link the references of pieces of equipment, their description, a high criticality and an absence of production shutdowns. When the rules with confidence = 1 expressing an equipment downtime are selected, 51 rules are found, concerning two devices: one allowing to store and distribute sand, and a mobile gateway. In the second case, it can be seen that the downtime of this strategic equipment has not required to stop production. The four corresponding rules, summarizing 369 transactions, nevertheless exhibit a recurrent problem that should be addressed.

Sem:

This rule base is interesting since its class diagram shows that it contains unusual details on the operator in charge of the maintenance activity. Moreover, the relatively limited number of rules allow a better legibility of the graphical representation.

Objective analysis: Rules with low support have widely dispersed confidence. On the contrary, few rules have high support and confidence. Especially, two symmetrical rules have again a maximum support and confidence (top-right corner of Fig. 8):

Maint_type:”Curative”= > Family_name=”DGF” sup = 0.96 conf = 1 lift = 1

Family_name=”DGF”= > Maint_type:”Curative” sup = 0.96 conf = 0.96 lift = 1

These rules mean that nearly all the transactions (sup = 0.96) concern curative activities on the system DGF, also concerned by some rare preventive maintenance activities (637 transactions on 14524).

Many rules of confidence 1 are accessible (top line of the graph of Fig. 8) linking for instance the category of the actor and the family of the system, or the family and the type of action. A good example is:

Maint_type=”Curative”. Actor_category=”Planner”= > Family_name=”DGF” sup = 0.20 conf = 1 lift = 1

showing that planners doing curative actions always intervene on the system family “DGF”.

Many rules with high confidence and lift are grouped in the points at the top of Fig. 9: coming back to the spreadsheet, it can be seen that 996 rules have a confidence = 1 and a lift = 1. They express strong but poorly informative relationships in the database, linked to empty fields: in a transaction, when a field is empty, many others are usually also empty.

Subjective analysis: The class diagram suggests interesting possible correlations, for instance between actor category and maintenance event type, duration and cause, duration and symptom, etc.

When selecting the filters, three categories of actors appear: planner (3104 transactions), technician (9177 transactions) and equipment owner (2238 transactions).

Table 3
Comparison of the results.

	Data model	Trans /rules	Genericity	Pos correlations
Aero	Basic	0,002	Varied	Low
Auto	Ambiguous	0,352	Poor	High
Pharm	Standard	0,005	Poor	High
Maint	Complex	0,020	Varied	Low
Sem	Standard	0,218	Poor	Average

Two types of activities (action and analysis) are distinguished in the data base. All actors are more oriented on actions, with typical rules like:

Actor_category="equipment_owner"= > Maint_event_type="Action" sup = 0.12 conf = 0.78 lift = 1.1

Actor_category="equipment_owner"= > Maint_event_type="Action" sup = 0.03 conf = 0.22 lift = 0.76

Some rules show what type of actor intervenes on each equipment. When linking symptoms to causes, it can for instance be seen that the same cause may sometimes have two different symptoms (which is consistent with the cardinality of Fig. 5). In the following rules, the second symptom has for instance a better confidence and correlation:

Maint_cause="Applicator" =>Maint_symptom="DO" sup= 0.02 conf = 0.37 lift = 4.48

Maint_cause="Applicator" = > Maint_symptom="RPS applicator" sup = 0.02 conf = 0.5 lift = 6.06

The main characteristics of each case study are summarized in Table 3, showing that generating rules from data bases is a subject on which it is difficult to draw simple conclusions. Good sense would suggest that more rules are likely to be generated if many transactions are provided but it is obviously false, since the number of generated rules depends on the number of frequent items, that depends on the data model. Nevertheless, this would require further investigation since there are no clear links between the two first columns of Table 3. The genericity of the rules is denoted by their support while the strength of the correlations between the IF and THEN parts of the rules is measured by their lift. As shown in Eq. (3), there is a link between support and lift, and rare combinations of events are more likely to generate high correlations: this can empirically be verified by seeing that the rule bases of poor genericity include event with the highest correlations.

5. Lessons learnt

These experiments have indeed confirmed our interest for "a posteriori" filtering: in many cases, the systematic evaluation of a large rule base allowed us not only to verify already known generic knowledge but also to find unexpected one, for instance linked to the way the data are entered (or not) by the operators. The examples given in previous section may seem to be anecdotal: we have not extracted from the databases critical knowledge that would completely change the point of view on the maintenance activities. Nevertheless, we managed to exhibit modest but precious areas of improvement, that went unnoticed until now because they were not on the mainstream information flow failure-cause-action that focuses all the attention in the maintenance process.

We have shown in Section 4 that the number of rules to be considered is not an obstacle for analysing the knowledge base if convenient tools are used. In that purpose, we have tried to show that combining in a flexible way objective assessment (based on the mapping of the rule base) and subjective assessment (based on class diagrams of the input data) may allow to explore different types of knowledge in a systematic way (generic knowledge and outliers; knowledge on the maintenance activities or on the links between concepts, etc.).

Another comment is that even if we have noticed that the data analyst can be of tremendous help during the analysis of the rules, their final interest can only be assessed by the users. In that purpose, a strong point of the suggested method is the use of a classical tool like a spreadsheet, allowing to transfer immediately the knowledge base to the users who, according to our experiments, can efficiently "play" with the rule base after a very limited training. Nevertheless, the final objective of the study was indeed not to transfer knowledge but a methodology and tools allowing the industrial partners to perform similar studies in the future without assistance. The use of a spreadsheet for exploring the rule base was a major point in that purpose. Similarly, we have noticed that a basic understanding of UML, obtained after half an hour of explanations, was enough for allowing maintenance engineers to be a force of proposal for the exploitation of the knowledge base. Nevertheless, we have also noticed that if a basic analysis of the rule base could be done autonomously by the maintenance engineers, a thorough analysis still requires some experience in data analysis.

Even if only knowledge on the maintenance domain was initially expected, it has been interesting to notice that many difficulties in data entry were detected, for instance through rules mentioning empty fields, that may represent a significant percentage of the produced knowledge. These difficulties may denote ambiguities in the data base structure or difficulties for the operator to find the required information. After further investigation, it seems that most of these ambiguities were in practice linked to attempts of the companies to simplify the data bases, resulting for instance in setting into question the clear distinction between failure, symptom and cause. In our opinion, managers, operators and data analysts should agree on a common interpretation of the attributes of a transaction before performing any change on the standard records.

As briefly illustrated in Section 4, we think that the interpretation of symmetrical rules (i.e. two rules in which IF and THEN parts are permuted) deserves specific attention even if, according to our experience, their interpretation remains difficult for end users. The absence of a rule may also be informative: indeed, when a bilateral relationship seems to exist between two attributes, this relationship should be expressed by two symmetrical rules, even if their confidence may be different. We have often checked that the absence of symmetrical rule led us to reconsider the strength of apparently well established relationships, between failures and symptoms for instance.

In any case, comparing what is expected according to the class diagrams and what is obtained in the rule bases may indeed help to check the consistence between the theoretical understanding of the relationships between attributes of the data base, denoted by the Class diagram, and the practices of the users.

Another point is that using a standard Apriori algorithm, it was only possible to generate rules with measures of interest exceeding chosen thresholds, preventing to have access to very specific rules. Since the first tests have shown the potential interest of rules of very low values of the measures of interest, we are now making some tests on a non-standard Apriori algorithm allowing to define thresholds as intervals. This should help us to have access to very specific rules, with the drawback of their possibly difficult interpretation.

6. Conclusion

Knowledge discovery from manufacturing databases is a very promising domain for industrial companies. The academic literature shows that many efficient algorithms are available in that purpose. Nevertheless, knowledge extraction, for instance by building rules, is not an end in itself. The analysis of the obtained knowledge base, sometimes neglected in the literature, is a mandatory step for improving the adoption of these techniques by companies.

Unlike what is often suggested, we have tried to show that generating first a huge mass of knowledge, then filtering this knowledge base textually using simple tools, may be an interesting alternative to an "a priori" filtering of the generated knowledge or to visual filtering. Our experiments have shown that generating many rules allows to have a better idea on the distribution of the measures of interest in the rule bases, that can help to assess to potential interest of a knowledge base. Similarly, building "maps" of the rule bases by correlating the measures of interest also helps to identify "zones" of rules of potential interest.

The empirical methodology described in this article will now be refined using other collected databases. A major challenge is now in the use of the "free texts" that should be of tremendous interest if correctly aggregated using text mining techniques. This should allow to go further in the understanding of the link existing between the data structure and the generated knowledge. Another point under investigation is the analysis of the evolution through time of the databases, that can be clearly indicated by the evolution of the generated rules. This could allow to analyse more objectively the effect of changes in the maintenance strategy.

At this step, even if we think that the suggested approach is promising, it still suffers from some limitations, the first one being that, as discussed in Section 5, the use of a conventional Apriori algorithm does not allow to have a complete view on the knowledge base that can be extracted from the database. In that purpose, it would be necessary to limit the number of generated rules while preserving their diversity. We have seen during our experiments that a lot of close rules are generated. We conduct now experiments in order to only generate some “representatives” of each kind of rule in order to generate a knowledge base easier manipulate, but representative from the whole potential knowledge base.

References

Accorsi, R , Manzini, R , Pascarella, P , Patella, M , & Sassi, S (2017) Data mining and machine learning for condition-based maintenance *Procedia Manufacturing, 11*, 1153–1161

Agrawal, R , Imielinski, T , & Swami, A (1993) Mining association rules between sets of items in large databases *Proceedings of the ACM SIGMOD international conference on management of data, SIGMOD 93* (pp 207–216) Washington, D C , USA: ACM May 26-28

Agrawal, R , & Srikant, R (1994) Fast algorithms for mining association rules in large databases *Proceedings of the 20th international conference on very large data bases, 12–15 September* (pp 487–499) Morgan Kaufmann Publishers Inc

Alsyouf, I (2009) Maintenance practices in Swedish industries: Survey results *International Journal of Production Economics, 121*(1), 212–223

Arif-Uz-Saman, K , Cholette, M E , Ma, L , & Karim, A (2016) Extracting failure time data from industrial maintenance records using text mining *Advanced Engineering Informatics, 33,* 388–396

Baohui, J , Yuxin, W , & Zheng-Qing, Y (2011) The research of data mining in AHM technology based on association rule *Proceedings of prognostics and system health management conference (PHM), IEEE, Shenzhen, 24-25 May* (pp 1–8)

Ben-Daya, M , Duffuaa, S O , Raouf, A , Knezevic, J , & Ait-Kadi, D (2009) *Handbook of maintenance management and engineering.* Springer

Bin, Z , & Wensheng, X (2015) *An improved algorithm for high speed train s maintenance data mining based on MapReduce. 25th international conference on could computing and big data, November 4-6, Shanghai, China*

Bruzzese, D , & Davino, C (2008) Visual mining of association rules In S J Simons (Ed) *Visual data mining, LNCS 4404* (pp 103–122) Berlin Heidelberg: Springer-Verlag

Choudhary, A K , Harding, J A , & Tiwari, M K (2009) Data mining in manufacturing: A review based on the kind of knowledge *Journal of Intelligent Manufacturing, 20*(5), 501–521

Crespo Marquez, A , & Gupta, J N D (2006) Contemporary maintenance management: Process, framework and supporting pillars *Omega - International Journal of Management Science, 34*(3), 313–326

Djatna, T , & Alitu, I M (2015) An application of association rule mining in total productive maintenance strategy: An analysis and modelling in wooden door manufacturing industry *Procedia Manufacturing, 4,* 336–343

EN13306:2001 (2001) Maintenance terminology - European standard CEN (European Committee for Standardization), Brussels

Fowler, M (2004) *UML distilled* (3rd ed) Addison Wesley

Garcia, E , Romero, C , Ventura, S , & de Castro, C (2009) An architecture for making recommendations to courseware authors using association rule mining and collaborative filtering *User Modeling and User-Adapted Interaction, 19*(1–2), 99–132

Gasmi, G , Ben Yahia, S , Mephu Nguifo, E , & Slimani, Y (2006) IGB: Une nouvelle base générique informative des règles d’association *Revue I3 (Information Interaction Intelligence), 6*(1), 31–67

Gasmi, G , BenYahia, S , Nguifo, E M , & Slimani, Y (2005) A new informative generic base of association rules *In proceedings of the ninth Pacific Asia knowledge discovery and data mining conference (PAKDD 05), Hanoi, Vietnam, May* (pp 81–90)

Geng, L , & Hamilton, H J (2006) Interestingness measures for data mining: A survey *ACM Computing Surveys, 38*(3) Article 9

Glavar, R , Kemeny, Z , Nemeth, T , Matyas, K , & Monostori, L (2016) A holistic approach foe quality oriented maintenance planning supported by data mining methods *Procedia CIRPP, 57*, 259–264

Grabot, B (2017) Fouille de données en maintenance: Exploitation des résultats, CIGI 2017, May 3-5, Compiègne, France

Hahsler, M , Chelluboina, S (2015) Visualizing association rules: Introduction to the R-extension package arules Viz Technical report, February

Han, J , & Kamber, M (2006) *Data mining: Concepts and techniques.* Morgan Kaufmann Publishers

Harding, J A , Shahbaz, M , Srinivas, & Kusiak, A (2006) Data mining in manufacturing: A review *Journal of Manufacturing Science and Engineering - Transactions of the ASME, 128*(4), 969–976

Karray, M H , Chebel-Morello, B , Zerhouni, N (2011) A formal ontology for industrial maintenance Terminology & ontology: Theories and applications, TOTh Conference 2011, May, Annecy, France

Köksal, G , Batmaz, I , & Testik, M C (2011) A review of data mining applications for quality improvement in manufacturing industry *Expert Systems with Applications, 38*(10), 13448–13467

Koskinen, K U (2012) Problem absorption as an organizational learning mechanism in project-based companies: Process thinking perspective *International Journal of Project Management, 30*(3), 308–316

Kotsiantis, S , & Kanepoulos, D (2006) Association rule mining: A recent overview *GESTS International Transactions on Computer Science and Engineering, 32*(1), 71–82

Larose, D T (2005) *Discovering knowledge in data: An introduction to data mining.* Wiley-Interscience

Liao, S H , Chu, P H , & Hsiao, P Y (2012) Data mining techniques and applications - A decade review from 2000 to 2011 *Expert Systems with Applications, 39*(12), 11303–11311

Mansingh, G , Osei-Bryson, K -M , & Reichgelt, H (2011) Using ontologies to facilitate post-processing of association rules by domain experts *Information Sciences, 181*, 419–434

Mansingh, G , Osei-Bryson, K M , Rao, L , & McNaughton, M (2016) *Data preparation: art or science? Proceedings of third IEEE international conference on data science and engineering ICDSE, 23-25 August, Kochi, India*

Maquee, A , Shojaie, A A , & Mosaddar, D (2012) Clustering and association rules in analyzing the efficiency of maintenance system of an urban bus network *International Journal of System Assurance Engineering and Management, 3*(3), 175–183

Palanisamy, S K (2006) *Association rule based classification.* Massachusetts, USA: Thesis from Worcester Polytechnic Institute

Potes Ruiz, P , Kamsu-Foguem, B , & Grabot, B (2014) Generating knowledge in maintenance from experience feedback, knowledge based systems, special issue on “Enhancing experience reuse and learning”, 68, pp 4–20

Ruschel, E , Alves Portela Santos, E , & de Freitas Rocha Loures, E (2017) Mining shop-floor data for preventive maintenance management: Integrating probabilistic and predictive models *Procedia Manufacturing, 11*, 1127–1134

Sammouri, W , Come, E , Oukhellou, L , Aknin, P , Fonlladosa, C -E , & Prendergast, K (2012) Temporal association rule mining for the preventive diagnosis of onboard subsystems within floating train data framework *Proceedings of the 15th international IEEE conference on intelligent transportation systems (ITSC), IEEE, Anchorage, 16–19 September* (pp 1351–1356)

Scheffer, T (2005) Finding association rules that trade support optimally against confidence *Intelligent Data Analysis, 9*(4), 381–395

Tan, P -N , Steinbach, M , & Kumar, V (2006) *Introduction to data mining.* Addison-Wesley

Yao, Y Y , Chen, Y H , & Yang, X D (2006) A measurement-theoretic foundation of rule interestingness evaluation In T Y Lin (Ed) *Foundations and novel approaches in data mining* (pp 41–59) Berlin: Springer-Verlag

Young, T , Fehskens, M , Pujara, P , Burger, M , & Edwards, G (2010) Utilizing data mining to influence maintenance actions *Proceedings of AUTOTESTCON systems readiness technology conference* (pp 267–271) Orlando: IEEE 13–16 September