Discovering Useful Knowledge from Aircraft Operation/Maintenance Data

Letourneau, Sylvain; Famili, Fazel; Matwin, S.

National Research Council Canada
Conseil national de recherches Canada

Institute for Information Technology
Institut de technologie de l'information

NRC · CNRC

Discovering Useful Knowledge from Aircraft Operation/Maintenance Data *

Letourneau, S., Famili, A., and Matwin, S.
July 1997

* published in the Proceedings of the Workshop on Machine Learning in the Real World, at the 14th International Conference on Machine Learning. London, England. July 8-12, 1997. pp. 34-41. NRC-40199.

Canada

Discovering Useful Knowledge from Aircraft Operation/Maintenance Data

Sylvain Létourneau
Institute for Information Technology
National Research Council Canada
Ottawa, Ontario, Canada, K1A 0R6
sletour@ai.iit.nrc.ca

A. Famili
Institute for Information Technology
National Research Council Canada
Ottawa, Ontario, Canada, K1A 0R6
famili@ai.iit.nrc.ca

Stan Matwin
Dept. of Computer Scienc
University of Ottawa
Ottawa, Ont. K1N6N5 Canada
stan@csi.uottawa.ca

Abstract

In this paper we present an overview of our research in discovering useful knowledge from data acquired during the operation and maintenance of a fleet of commercial aircraft. In particular, we discuss the application domain and explain some of the constraints that we have encountered in analyzing this data. We present the approach that we have developed to efficiently analyze large amounts of diverse forms of data available in this domain. Preliminary results for one of the problems that we have investigated are given. Further work and challenging issues are proposed at the end.

Keywords: Knowledge Discovery, Data Mining, Feature Engineering, Intelligent Decision Support, Aircraft Operation

1 INTRODUCTION

Today, in almost all industries, vast amounts of data, numeric and symbolic, are continuously generated. This is due to increasing progress in the development of process monitoring and data acquisition systems. The situation is particularly true in the operation/maintenance of commercial aircraft where large number of on-board sensors automatically measure status of the aircraft components and their operation along with conditions surrounding the aircraft. Almost all these data are transmitted to a central database management system where they are preprocessed and stored in a large database. In most cases, these data may not be used, or even properly warehoused. Several reasons exist: (i) engineers and operators do not have sufficient time to analyse terabytes of data, (ii) complexity of the data analysis process is sometimes beyond the simple application of a data analysis tool (machine learning or other), and (iii) there is no well defined automated mechanism to extract, preprocess and analyze the data and summarize the results so that the engineers and technicians can use it. On the other hands, it is obvious that valuable information and scientific results can be obtained from an appropriate use of this data.

A knowledge discovery application that discovers valuable patterns from the operation/maintenance data can be very beneficial. Due to large amounts of investments by airlines and the high level of safety, any discovered patterns in the data that predict or explain component failures may lead to saving of several thousands of dollars, reduce the number of delays, increase the overall level of safety, and help to get a more in-depth understanding of the complex systems involved. Considering the large amounts of data systematically collected and the availability of domain knowledge, there are good reasons to believe that useful information can be discovered from this scientific domain [Fayyad et al. 1996a]. However, the complexity of this application (diverse forms of data, time series relationships, high dimentionality, imbalance number of positive and negative examples [Kubat and Matwin 1997], presence of contexts [Turney 1996]) makes development of an appropriate knowledge discovery strategy difficult.

In this paper, we discuss the specific issues to consider during the analysis of commercial aircraft data. We further introduce a knowledge discovery in databases (KDD) approach that we are developing to discover hidden information from this data. The approach is presented as a sequence of four processes that can be followed to investi-

Workshop on Machine Learning Applications in the Real World: Methodological Aspects and Implications. ICML-97.

gate problems of interest. For each process, we explain the goals, the techniques that can be used, the domain information available and the expected results. We emphasize on processes which are specially problematic in our application and therefore, require more research work. We argue that the difficulties identified are generally important since they can be encountered in other real world applications as well. The overall iterative and interactive nature of the KDD process [Fayyad et al. 1996b] is supported by our approach.

The paper is organized as follows. We first elaborate the problem and the main characteristics of the data. Then, we explain the knowledge discovery approach and the steps to be taken. In section 4, given a specific problem of interest, we present the use of the proposed approach and report preliminary results. We conclude the paper in section 5 and list a number of challenges that we see ahead of us.

2 STATEMENT OF THE PROBLEM

A large Canadian airline, a major builder of aircraft engines, and the National Research Council of Canada are collaborating to develop an Intergratred Diagnosis System (IDS)[Wylie et al. 1997]. This system will provide support to the airline maintenance staff to accurately predict problems, to obtain an explanation for engine performance deviations, and to monitor the overall status of the aircraft in real time. The software can therefore be used to improve the decision making process of the aircraft maintenance. As part of the knowledge acquisition process, we aim at extracting useful information from large amounts of data collected since fall 1994. These data contain the information on all 34 Airbus A320 aircraft of the airline. The information recorded includes the sensor measurements taken on each individual aircraft, documentation of aircraft problems along with the operation/maintenance actions taken for each of them, and warning and failure messages that are automatically generated by the aircraft on-board computers when particular conditions occur. As an indication of the size of the data available, each A-320 produces about 1.5 Megabytes of data per month.

In this paper, we focus on an approach that we are developing to extract useful information from these data. After a sucessful evaluation and validation by domain experts, the discovered knowledge will be incorporated into the IDS. IDS supports two types of reasoning: cases-based and rule-based. We are therefore specially interested by knowledge that can be mapped into one of these representations.

The first major challenge in our project has been warehousing all these data that we receive so that: (i) the contents of database are clean, (ii) no useful information is lost, (iii) required data can be retrieved and used as efficiently as possible, (iv) no irrelevant and redundant data are stored anywhere, and (v) all forms of reasoning can be performed with minimum data extraction and preprocessing efforts. Our application has involved all five challenges of data warehousing, mentioned above.

The following listing shows some of the important issues in this application:

Data format

The data comes in different forms and is scattered in various databases that contain different data structures. There are three types of data. First, the reports generated by the aircraft on-board monitoring system are composed of numeric and symbolic parameters values. Second, the warning and failure messages are in fixed textual format. Finally, the descriptions of aircraft operation/maintenance problems, called snags reports, are provided in free textual forms with many inconsistent abbreviations.

Data quality

Like any other real world applications, we noticed several problems with the data. These were: missing parameter values, improper data types, out-of-range data, incomplete records or instances, and unavailable data.

Data complexity

The overall data characteristics is fairly complex. In addition to multiple sources of information, data comes in several levels of granularity (e.g. fleet, aircraft, engine, engine report, duration). The data dimensions are high (i.e. number of parameters and number of records per report or per level of granularity). Several parameters are expected to have time series relationships. In some cases, we have very large data sets that contain only a small number of positive examples.

Domain information

There are various forms of background knowledge that are available in the form of on-line and hard copy documentation that have been written by different manufacturers and the airline. Examples are troubleshooting guides, training manuals, empirical studies, etc. Proper use of this background knowledge at different stages of data preprocessing and data analysis is crucial. For example, identifying all classes or problem names and their definitions, selecting

relevant parameters for labelling each instance, identifying out-of-range thresholds, are the type of information that we can obtain from the background knowledge.

Presence of contexts

Some sensor measurements can be influenced by contextual conditions. For example, the measurements of the `exhaust-gas-temperature` of an engine may be influenced by: the altitude of the aircraft, the actual outside temperature, and the age of the aircraft. To obtain meaningful results, this parameter should be normalized. The difficulty comes from the fact that we generally do not know what the required transformations are. Inferring appropriate normalization formulas from the data represent a difficult challenge that has been recently addressed by other researchers [Katz et al. 1990; Turney 1996].

Given the above characteristics, our overall goal is to develop a methodology that can be used to either explain component failures/performance deviations or help in predicting the occurrence of future problems. We define *component failures* as problems in which a particular component or subsystem fails without a known reason. Examples are temperature sensor at fault or auxiliary power unit control computer at fault. *Performance failures* (or performance deviations) represent conditions upon which one or several performance parameters deviate from their usual ranges. Examples are when the engine exhaust gas temperature is above an expected limit or when engine shaft speed is below a limit at the peak fuel flow. We now turn to the methodology that we are developing to address such problems.

3 THE KNOWLEDGE DISCOVERY APPROACH

Given the application characteristics presented in the previous section, our goal is to develop an overall data analysis methodology that can be applied to find patterns which explain or predict component/performance failures. We assume that the data warehousing tasks have already been done. Four processes compose the approach (see Figure 1). Each process generates output that can be used in further steps. The right part of Figure 1 shows the expected output for each process. The overall approach is driven by an investigation description. An investigation description is simply a question that expresses a problem of interest in the given domain. Examples of investigation descriptions for the analysis of the auxiliary power unit (APU) engine of the Airbus A-320 are:

1. Can we predict an APU starter failure problem?
2. Can we build a model to assess the overall health of an APU?
3. Can we come up with explanations for "unexpected" exhaust gas temperature at peak of the APU shaft speed (i.e. a performance failure)?

The first two invesatigation questions are understood as component failures while the third one is undestood as a performance failure. The idea of starting the overall KDD process by a problem definition is not new. Most of all statistical inference techniques start by a problem definition referred as *hypothesis*. Recent work in machine learning has also pointed out the importance of guiding the analysis toward a specific problem definition [Saitta et al. 1995]. The question form seems to be appropriate to formulate the investigation descriptions since it is usually precise and easy to understand by both the domain expert and the analysts. It is important to note that an investigation question cannot simply be answered by a positive affirmation. To be complete and acceptable, a positive answer to an investigation question must be supported by a concrete solution (or model) and its empirical evaluation. On the other hand, a negative answer is usually enough since it is generally impossible to prove the non existence of a solution by using the available data only.

The goal of the application of the discovery process is to answer a given investigation description. The tasks and techniques involved in the aircraft domain are explained below.

3.1 IDENTIFY THE RELEVANT DATA SOURCES

The first step of the approach consists of identifying the sources of information that are related to the selected investigation question. This step is required in our application because many kinds of data are available and it is not always obvious to determine which one of them can be useful for the current investigation problem. For examples, the on-board computers of the A-320 generate up to 11 types of reports only to describe the engine operation status under different conditions. Each report typically contains between 90 and 150 parameter values (about 2/3 of them are numerical values). Since each report has its own structure, one cannot easily analyse data from different reports at the same time. One possibility is to focus on the reports that are pertinent to the current investigation problem.

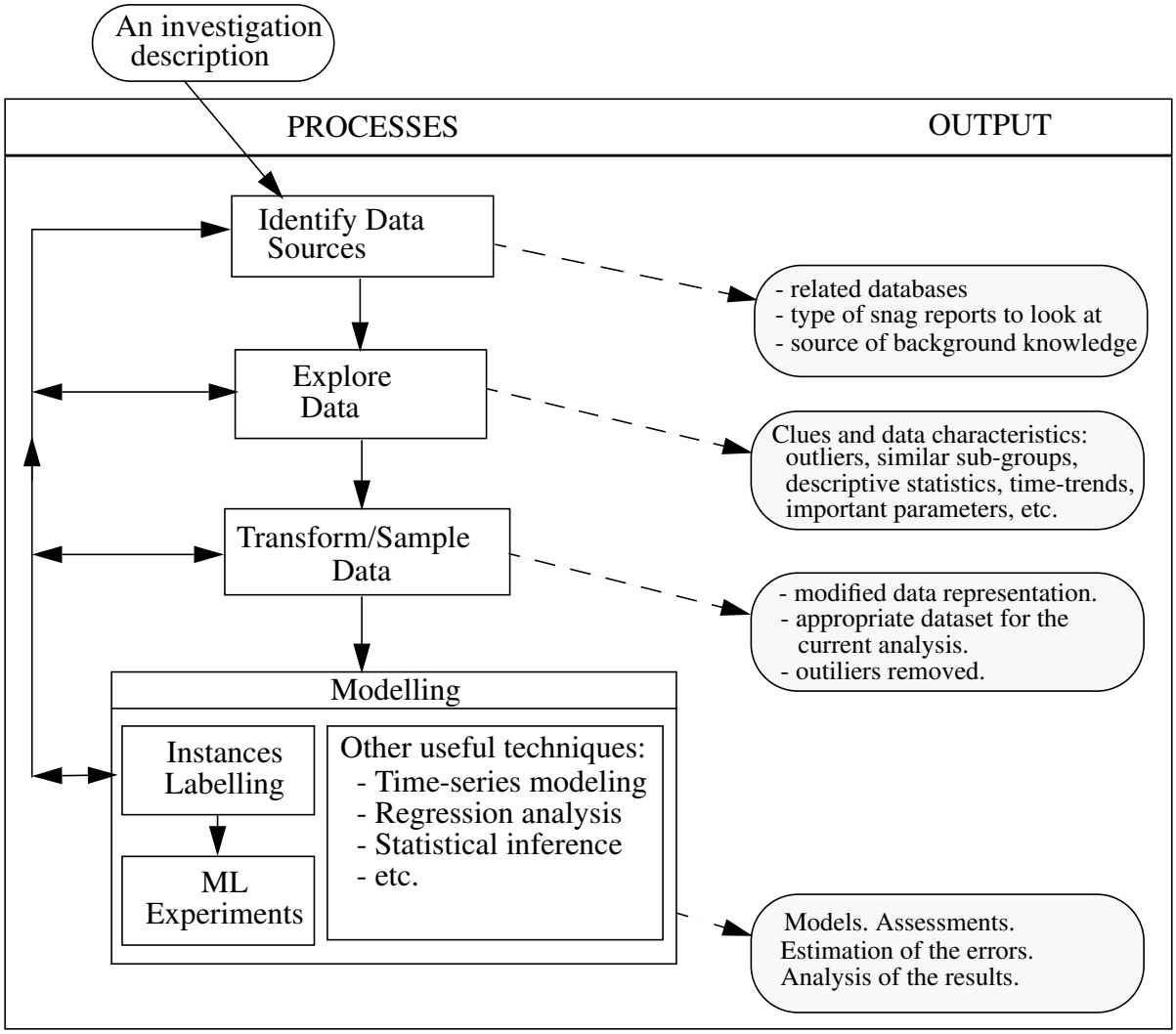

Figure 1: The main processes of the discovery approach.

These reports constitute the primary source of information for the analysis. However, important clues can be found from other sources of information such as the descriptions of aircraft problems and repair actions and the list of warning and failure massages. The information contained in these two latter sources of information may help to reduce the size of the data by providing lead to the most important subsets of data. Note that the search in the descriptions of aircraft problems and repair actions is difficult to automate since these do not follow a consistent format. A given problem may be described in a different way from one report to another. Domain experts and technical documentation (e.g troubleshooting manuals) can help to select the most important parameters and therefore reduce the search space. These are examples of background information in the aircraft domain that can make the difference in the overall analysis.

At the end of this first process, the analyst should know what are the relevant subsets of data for the current investigation problem and may have an idea about the important parameters.

3.2 EXPLORE SELECTED DATA

The second step consists of an in-depth exploration of the selected sources of information. The goal is to find the main characteristics of the data related to the current investigation. For example, if one is interested in analysing the starter failure problem, the exploratory analysis

should include: a search for the number of aircraft that had this problem in the past, identification of the parameters that seem to be related to this problem, search for time-series trends around the occurrences of the problem, compute descriptive statistics for the important parameters (e.g. average, variance, coefficients of correlation or auto-correlation in the case of time-series patterns), and study of the underlying distribution of the data. Statistical exploratory methods and visualization techniques are mainly used to collect these valuable pieces of information. The information collected during this step can definitely influence the subsequent steps of the analysis. For instance, if time series trends are observed during this step, then the data will have to be transformed (in the next process) into a more suitable format for classification.

Note that the exploratory process described here do not represent an exhaustive approach. We are focusing on information that may help to solve a specific investigation problem only. The search space is limited to the subsets of data that have been selected as relevant in the first process.

3.3 TRANSFORM AND SAMPLE THE DATA

The third process is composed of two tasks: data transformation and sampling. The goal of this process is to build a dataset that will be appropriate for modelling (last step). In our specific domain, a lot of work is typically required on the representation of the data. We first need to identify and remove irrelevant features from the data since they can affect the quality of the final results. Automatic feature subset selection approaches (e.g. the wrapper model [John et al. 1994] or filter model [Almuallim and Dietterich 1992]) are not suitable for our application for two reasons: i) we have to much data to work with them and ii) the available methods are not able to handle domain specific information. Secondly, as explained in Section 2, we need to normalize some sensor measurement values according to external conditions (e.g. temperature, pressure, altitude). This process may be quite complex since the normalization formulas are usually unknown. Identifying the contextual features from the data is still a research challenge [Turney 1996]. Finally, we may also want to improve our representation by extracting features from time-series patterns that seem relevant to the current investigation. Many function approximation techniques can be used to create the new features (e.g. fast fourier transform, k^{th}-moment procedure). Experimentation with different feature extraction techniques may be necessary to find the most appropriate features for the given time-series patterns.

According to the information acquired from the exploratory analysis, one may decide to only select subsets of the data (i.e. sampling the data). An obvious reason to limit the analysis to subsets of data comes from the volume of available data: there is simply too much data (several thousands of records with about 100 parameters in each). Thesecond reason is related to the meaning of the end results. In order to obtain meaningful models, only the relevant data for the selected problem have to be used. For example, with the starting prediction problem, the only data that are pertinent to build the model are those which preceded the occurrences of the problem. There is no need to include thousands of records from the aircraft which never had this problem. A simple random sampling strategy, such as proposed in [Fayyad et al. 1996b] for an image classification problem, is not appropriate for our application, since more meaningful data can be selected by using background knowledge and clues from the exploratory analysis.

3.4 BUILD MODELS

The last process is called modelling. The goal of this process is to build one or several models to answer the investigation question. Depending on the information collected during the previous processes and the type of results desired, machine learning techniques such as decision tree induction, rule induction, instance-based, or neural networks can be considered to build models. A variety of statistical techniques may also be used, specially during the evaluation of the results.

As indicated in Figure 1, an additional step, called labelling, is required before conducting a machine learning analysis. The labelling task consists of assigning each example to one of the pre-defined categories (also referred as class). In "traditional" machine learning applications the examples are pre-classified by a teacher (domain expert) and the analysts do not have to get involved in labelling. However, in a real-world application like the one described here, the volume of data is so large that no human can label each example manually. An automatic or semi-automatic procedure is therefore required to label the instance. We see this problem as an obstacle that must be addressed carefully for successful application of machine learning methods in real world applications. Two related reasons are: i) the applicability and usefulness of the learned models directly depends on the quality of the

labelling, and ii) the labelling procedure may be very difficult to automate since it typically involves considerable amount of expertise in the target domain. On the other hand, the complexity of the labelling process may depend on the sampling strategy. More research work on sampling strategies for large data sets [Musick, Catlett, and Russel 1993] is therefore also important.

As shown in Figure 1, the overall approach is iterative. In any cases, one may decide to repeat some previous steps. In fact, several iterations are often required to obtain interesting results. Referring to Figure 1, it is important to note that the results obtained from one process are not only used by the following process but by any of the successive processes. For example, results from the exploratory analysis can be used in the sampling as well as in the modelling process. Moreover, results obtained during the analysis of one investigation can also be used to study another investigation problem.

4 PRELIMINARY RESULTS

We now present the use of the data mining approach introduced in the previous section and report some results for the first investigation definition introduced in Section 3: "Can we predict an APU starter failure?"

The first process of the methodology lead us to three sources of information for this problem: textual description of the repairs for the APU, APU sensor measurements data set which contain about 90 parameters per record, and some technical documents about the APU. In the second process, the exploration of the data, we first searched trough the descriptions of the repairs and realized that we have only four documented occurrences of the investigated problem (APU starter failure) for the period of time considered. Always during the exploratory step, we visualized the APU sensor measurement parameters individually and observed two time-series patterns that looked relevant for the APU starter problem. These patterns were related to parameters STA (start time) and NPA (shaft speed) respectively. As shown in Figure 2, for a period of about one month before the failure of the starter, STA showed an increase while NPA decrease. Discussions with an expert in the domain confirmed the relevance of these parameters.

With the help of the above information, we decided to build a regression model that will characterize the behaviours of STA and NPA before the failure of the starter. Such a model could then be used in a monitoring system to predict APU starter failures. According to the lenght of the trends observed, the data selected (in the sampling process) corresponded to the periods of 30 days preceding the starter failures. We used a part of this data to create regression models for STA and NPA and tested them in the other part of the data. It turned out that with a very high confidence level (over 99%), the regression models were appropriate on the testing set. Domain experts acknowledged the relevance of the results in terms of understandability and usefulness for prediction.

However, we realized that a more exhaustive evaluation of the results is required. For this reason, we are repeating the first process of the proposed methodology (the identification of sources of information) to identify additionnal relevant datasets for the current investigation problem. Iteration trough the modelling process will be done as well to see if we still can improve the obtained results. In particular, we are trying to develop a labelling strategy that will allow us to use a supervised classification approach instead of a non-linear regression procedure in the modelling process.

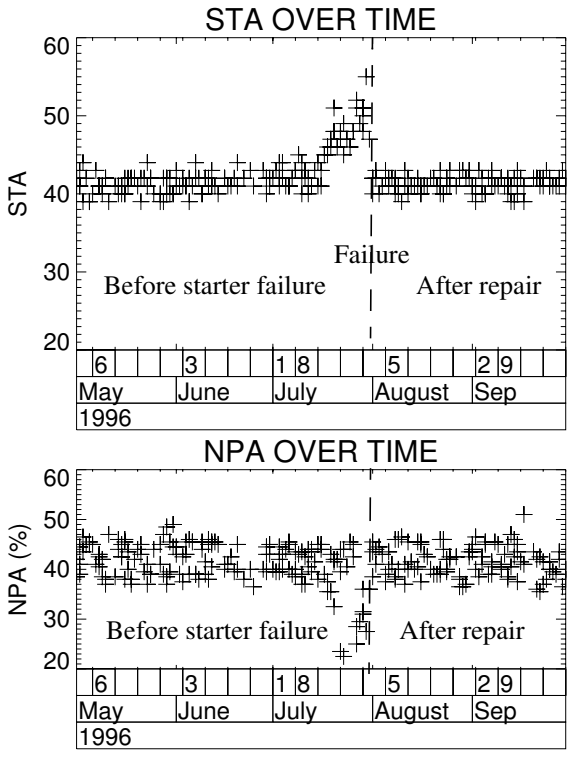

Figure 2. Typical NPA and STA patterns before the occurence of a starter failure.

5 CONCLUDING REMARKS AND CHALLENGES

In this paper we have discussed an approach for discovering some useful knowledge from large amounts of data that are manually and automatically generated during maintenance and operation of commercial aircraft. We have discussed several important issues in analyzing data in this domain. These issues were related to: data format, data complexity, domain information, and presence of contexts. We introduced a knowledge discovery approach that we have developed for this real world application. This approach consists of four steps: (i) identification of the relevant sources of information, (ii) exploration of the selected relevant data, (iii) sampling and data transformation, and (iv) modelling. All of these steps are guided by a specific investigation problem which has to be formulated with the help of domain experts before starting the analysis. The proposed approach helps to guide the analysis through the application of diverse discovery techniques. Such a methodological procedure will help us to address the complexity of the domain considered and therefore optimized our chance to discover valuable information. We presented preliminary results that plausibly confirmed this hypothesis but more experiments are clearly required.

During the presentation of the approach, we raised several difficulties that have to be addressed to successfully apply machine learning algorithms in complex real world domains. Among others, we noted the problems related to: the labelling of the instances, the selection of the relevant data, and the use of contextual information. In a long term project, such as the one described here, it may also be very important to address the following three issues: (i) finding an automatic approach to define relevant investigation problems for this domain, (ii) developing tools that are necessary to disseminate discovered knowledge to the end users, and (iii) automating most of the tasks involved in data mining process.

Acknowledgments

Thanks to Chris Sowerby and Francis Ruest from Air Canada to provide domain information and helped us during the evaluation of the results. Thanks to three anonymous referees of the Workshop for their comments on an earlier version of this paper.

References

Almuallim, H. & Dietterich, T. G. (1992). Efficient algorithms for identifying relevant features. In *Proceedings of the 9th Canadian Conference on Artificial Intelligence.* 38-45. Vancouver, BC: Morgan Kaufmann.

Brodley, C. & Smyth, P. (1996). Applying Classification Algorithms in Practice, *Statistics and Computing*, in press.

Fayyad, U., Haussler, D. & Stolorz, P. (1996a). KDD for Science Data Analysis: Issues and Exemples. In E. Simoudis, J. Han & U. Fayyad (eds.), *The Second International Conference on Knowledge Discovery & Data Mining, 50-56.* CA, AAAI Press.

Fayyad, U., Piatetsky-Shapiro, G. & Smyth, P. (1996b). *Advances in Knowledge Discovery and Data Mining*, AAAI/MIT Press.

John, G.H., Kohavi, R. & Pfleger, K. (1994). Irrelevant features and the subset selection problem. In *Proceeding of The 11th International Conference on Machine Learning.* 121-129. CA: Morgan Kaufmann.

Katz, A.J., Gately, M.T. & Collins, D.R. (1990). Robust classifiers without robust features, *Neural Computation*, 2, 472-479. Cambridge, Mass. : MIT Press.

Kubat, M. & Matwin, S. (1997). Addressing the Curse of Imbalanced Training Sets: One-Sided Selection. To appear in *Proceeding of The 14th International Conference on Machine Learning*. Nashville, Tennessee.

Musick, R., Catlett, J. & Russell, S. (1993). Decision Theoretic Subsampling for Induction on Large Databases, *Proceedings of the 10th International Conference on Machine Learning*, pp. 212-219. CA: Morgan Kaufmann.

Saitta, L., Giordana, A. & Neri, F. (1995). What is the Real World? In *Proceedings of the Workshop on Applying Machine Learning in Practice*, at the *12th International Conference on Machine Learning*. 34-40. CA.

Turney, P. (1996). The Identification of Context-Sensitive Features: A Formal Definition of Context for Concept Learning. *Proceedings of the Workshop on Learning in Context-Sensitive Domains*, at the *13th International Conference on Machine Learning*. 53-59. Bari, Italy.

Wylie, R, Ochard, R., Halasz, M. & Dubé, F. (1997). IDS: Improving Aircraft Fleet Maintenance. To appear in *Inovative Applications of Artificial Intelligence.* Providence, RI.