Chapter 1

Introduction

1.1 Knowledge Discovery in Databases

Knowledge Discovery in Databases - or *KDD* - is a term used to describe the procedure of acquiring high-level knowledge from low-level data. As a formal definition, *Knowledge Discovery* is the nontrivial extraction of implicit, previously unknown, and potentially useful information from data [2]. This knowledge is usually found in the form of patterns and relations between variables which were unlikely to be related.

The KDD process involves several steps [1] which can be summarized as follows:

1. **Understanding the problem:** The first step involves understanding the environment we are studying and gaining relevant prior knowledge. In this step we must identify which goals we want to set for the knowledge discovery process. This is, we must identify the kind of knowledge we want to obtain and the data we can count on for this process.

2. **Creating a target dataset:** We will usually need to select a subset of variables from the available datasets. While the system we are studying may need a lot of variables to log events or make data relations, we will not likely need all of them to characterize our problem. Reducing the dimensionality of the problem will provide better results and ease the following steps.

3. **Data cleaning and preprocessing:** In this step we have to discern which data is actually relevant and significant for our study, and which is merely noise or outliers which should be disregarded. Operations such as noise modelling or mapping of missing and unknown values are also taken in this step.

4. **Data mining:** For this step we must first have decided the purpose of the model derived by the data mining algorithm. For example, summarization, regression, clustering and others. According to our decision, some data mining algorithms will be more appropriate than others.

5. **Interpretation of results:** Consists on interpreting the discovered patterns, removing those redundant or irrelevant and translating the useful ones into understandable terms.

6. **Consolidation of discovered knowledge:** The discovered knowledge is finally consolidated in an appropriate form. Depending on the context of our project, it might be simply documented or integrated in predictive modules for the analysed systems.

1.1.1 Understanding the problem

In our project, the application domain is railway maintenance systems. Although it is not necessary for us to know all the details about railway systems or their maintenance, we do need to have at least general knowledge about the kind of systems we are going to work with. Specifically, it is of interest to us to know about the kind of elements from which we will obtain information, and how they are related to each others. Furthermore, it is of special need for us to understand how the maintenance systems record information about raised alarms. A deep insight into the provided databases is essential, as we cannot work with any data unless we have idea of what does it represent.

Definition of objectives

The main goal of the knowledge discovery process in this project will be achieving the ability of predicting future alarms based on present and past logged events.

The maintenance stations of the studied network are able to raise alarms automatically when certain events happen. These alarms are shown to operators and classified according to their severity. The specific main goal of our project is to be able to show also which alarms are likely to be raised in the short future, based on which alarms have been being raised previously.

Using this method, we can achieve knowledge such as the following:

- Obtain sequential associative rules. These indicate which events are going to happen in the future according to which events have already happened, based on patterns found in the databases. E.g. *If Event 2 happens after Event 1, Event 3 is going to happen in the future.*

- Associate specific patterns to specific systems. For instance, not all the maintenance stations are of the same characteristics, as well as not all the railway lines are. This would allow filtering predictions depending on the type of station.

- Determine the precision (confidence) for each acquired prediction rule. This would allow users to filter predicted events for their precision and prioritise those events with higher possibilities to happen.

Furthermore, additional patterns and predictive techniques can be found using additional data, such as measures obtained on preventive maintenance procedures on several system elements. These measures, such as power supply levels or system temperatures, are potentially highly related to system failures, thus being possible to predict failures by monitorising said variables. However,

automation of the acquisition of these variables can be an additional problem which would need to rely on advanced systems which are currently out of our reach.

Alarm prediction based on said measures will therefore not be amongst our priorities, due to the additional problems we would have to face. In any case, we think this information can also be of enormous interest for failure prediction, even if its implementation is not as immediate as for alarm-based prediction.

Chapter 2

Data Mining Algorithms

2.1 Choosing an appropriate Data Mining algorithm

Data Mining comprises a large amount of different algorithms which can be used for the Knowledge Discovery process. Depending on the nature of the data on which we will apply these algorithms, and on the kind of knowledge we expect or want to acquire, we will need algorithms of different types.

Different algorithms can usually be classified in the following categories:

1. **Classification:** Learning a function that maps an item into predefined classes.
2. **Regression:** Learning a function that maps an item to a predicted variable.
3. **Segmentation:** Identifying a set of clusters to categorise the data.
4. **Summarization:** Finding a compact description for the data.
5. **Association:** Finding significant dependencies between different variables.[**?**]
6. **Sequence analysis:** Finding frequent sequences or episodes in data.[**?**][**?**]

The most immediate type of technique to be applied is *Sequence analysis*. When dealing with event-based problems, such as this one, it is very important to maintain the information given by the time variable. Alarms (and therefore their related events) are most likely to happen in sequential patterns. That is, the information about the order and timing of the alarms is very likely to be essential for the prediction tasks.

There are also techniques based on time constraints[**?**]. These techniques convert the event-based data into tuples with quantitative variables. For example, by sliding a temporal window over the data, we can convert the occurrence of events to numeric values, and thus apply other kind of techniques such as association or regression algorithms. This might come in specially handy when

taking into consideration additional variables such as the aforementioned measurements. In this case, we may disregard sequential information and focus only in frequency information, finding out how these frequencies relate to variables such as temperature or other environment data. There are four time constraints: sliding window, minimum time gap, maximum time gap and duration. A large amount of existing algorithms [**?**] have been implemented taking advantage of them, and can be highly useful for our project.

Bibliography

[1] UM Feyyad. Data mining and knowledge discovery: Making sense out of data. *IEEE expert*, 1996.

[2] WJ Frawley. Knowledge discovery in databases: An overview. *Ai Magazine*, 1992.

[]