Chapter XVIII
Knowledge Through Evolution

Russell Beale
University of Birmingham, UK

Andy Pryke
University of Birmingham, UK

ABSTRACT

This chapter argues that a knowledge discovery system should be interactive, should utilise the best in artificial intelligence (AI), evolutionary, and statistical techniques in deriving results, but should be able to trade accuracy for understanding. Further, it needs to provide a means for users to indicate what exactly constitutes "interesting", as well as understanding suggestions output by the computer. One such system is `Haiku`, which combines interactive 3D dynamic visualization and genetic algorithm techniques, and enables users to visually explore features and evaluate explanations generated by the system. Three case studies are described which illustrate the effectiveness of the `Haiku` system, these being Australian credit card data, Boston area housing data, and company telecommunications network call patterns. We conclude that a combination of intuitive and knowledge-driven exploration, together with conventional machine learning algorithms, offers a much richer environment, which in turn can lead to a deeper understanding of the domain under study.

INTRODUCTION

In this modern world, information is collected all the time: from our shopping habits to web browsing behaviours, from the calls between businesses to the medical records of individuals, data is acquired, stored, and gradually linked together. In this morass of data, there are many relationships that are not down to chance, but transforming data into information is not a trivial task. Data is obtained from observation and measurement, and has no intrinsic value. But from it we can create information: theories and relationships that describe the relationships between observations. And from information we can create knowledge: high-level descriptions of what and why, explaining and understanding the fundamental data observations. The mass of data available allows us to potentially discover important relationships between things, but the sheer volume dictates that we need to use the number-crunching power of computers to assist us with this process.

Data mining, or knowledge discovery as it is sometimes called, is the application of artificial intelligence and statistical analysis techniques to data in order to uncover information. Given a number of large datasets, we are fundamentally interested in finding and identifying interesting relationships between different items of data. This may be to identify purchasing patterns, which are then used for commercial gain through guiding effective promotions, or to identify links between environmental influences and medical problems, allowing better public health information and action. We may be trying to identify the effects of poverty, or to understand why radio-frequency observations of certain stars fluctuate regularly. Whatever the domain of the data, we are engaged in a search for knowledge, and are looking for interesting patterns in the data.

But what is "interesting"? One day, it may be that the data falls into a general trend; the next it may be the few outliers that are the fascinating ones. Interest, like beauty, is in the eye of the beholder. For this reason, we cannot leave the search for knowledge to computers alone. We have to be able to guide them as to what it is we are looking for, which areas to focus their phenomenal computing power on. In order for data mining to be generically useful to us, it must therefore have some way in which we can indicate what is interesting and what is not, and for that to be dynamic and changeable. Many data mining systems do not offer this flexibility in approach: they are one-shot systems, using their inbuilt techniques to theorise and analyse data, but they address it blindly, as they are unable to incorporate domain knowledge or insights into what is being looked for; they have only one perspective on what is interesting, and report only on data that fit such a view. Many such systems have been utilised effectively, but we believe that there is more to data mining than grabbing just the choicest, most obvious nuggets.

There are further issues with current approaches to data mining, in that the answers are often almost as incomprehensible as the raw data. It may be that rules can be found to classify data correctly into different categories, but if the rules to do so are pages long, then only the machine can do the classification: we may know how to do the classification, but have no insight into why it may be like that. We have gained information, but not knowledge. We believe that we should be able to understand the answers that the system gives us. In order to achieve this, it may be that we need broader, less accurate generalisations that are comprehensible to the human mind, but then feel confident in the main principles to allow the machine to do classification based on much more complex rules that are refinements of these basic principles. For example, "if it's red and squishy, it's a strawberry" is easy to understand. Even if that's true only 80% of the time, it's a useful rule, and easier to grasp than:

```
red, deforms 4mm under 2N pressure,
>3cm diameter = strawberry &
red, deforms 1mm under 2N pressure,
<6cm diameter = cherry &
red, deforms 3 mm under 4N pressure,
>5cm diameter = plum
else raspberry
```

which may be 96% correct but is hardly memorable. For many data mining systems, the rules developed are far more complex than this, each having numerous terms, with no overall picture able to emerge. For statistical-based systems, the parameter sets are even harder to interpret.

Since "interesting" is essentially a human construct, we argue that we need a human in the data mining loop; if we are to develop an effective system, we need to allow them to understand and