C43 Pros and Cons of Data Mining in the Forensic Context

Annabel Bolck, PhD, and Chrissie Schapers, Netherlands Forensic Institute, Laan van Ypenburg 6, Den Haag, 2497 GB, NETHERLANDS*

After this presentation, attendees will have a better understanding of data mining techniques.

This presentation will impact forensic science by showing the advan- tages and disadvantages of the use of data mining techniques in general and in the forensic context.

Data mining is a general term, which describes a number of techniques to identify pieces of information or decision-making knowledge in data. In contrast to many other techniques, data mining techniques do not work with a research hypothesis. This is both the strength and weakness of data mining. The techniques can help identify (seemingly useless) patterns in data that otherwise would not be discovered. Sometimes they can help to translate these patterns in valuable information. Other times it will result in drown- ing in the data.

Another strength of data mining is that it can handle huge amounts of data in automatic or semi-automatic way with the use of machine learning and computers. Though the techniques can also be applied to relatively small datasets (we will provide an example of that) it is especially designed for large datasets.

In this presentation a short (and not complete) overview is given of the use of data mining in forensics. The advantages and disadvantages of data mining in this context will be discussed in general and when applied to a few examples.

One of the examples concerns the clustering of MDMA tablets. 3,4- methylenedioxymethamphetamine (MDMA), the active compound of ec- stasy, is one of the most widely spread illegal synthetic drugs in Europe. The Netherlands is one of the most producing countries. Therefore the Netherlands Forensic Institute is involved in a large research project concerning ecstasy. One of the goals of this project is to investigate whether it is possible to cluster XTC consignments by production batch, producer, or production method based on the chemical composition of the tablets. Mainly unsupervised clustering techniques are used for this goal, because in most cases tablets are not found in production facilities, and therefore there true origin is not known. It is found that (unsupervised) clustering by production method is possible, but clustering by producer or production method is much harder, not in the last place as a result of the dynamic market of ecstasy producers.

Another example will concern profiling and finding indicators for offenders for specific crimes by the Dutch Police. The ultimate goal of such profiling is to increase the success and efficiency of police actions by mak- ing better informed decisions. The first challenge lies in retrieving the data and merging data sources in different formats together. Importantly, there are legal limitations with respect to analysis of crime records, especially in rela- tion to open source data. Finally, results that are extracted from the data need to be presented to the police in an understandable way on a routine basis.

Data Mining, Unsupervised Clustering, Drugs

 ** Presenting Author*