

“KDD: Introduction and Experiences”

Dr. Eduardo Morales

Review by **Roberto Hoyos Morales**

March the 31st, 2005

1. Main Ideas

Advances in data collection and process automation have generated a huge amount of data. The main problem derived is that this technological growth in data storing has not been matched with a correspondent growth in data analysis.

So the main motivation is to extract valuable information (translated in economical benefits) from the data. This requires the identification of novelty patterns, potentially useful for a business or enterprise.

2. Results and Conclusions by the speaker

The lessons from KDD Process in the following structure are as follow.

Selection.- We have to be very careful when selecting attributes and data, because sometimes ‘irrelevant’ attributes at the beginning may prove to be useful when obtaining new relations. At the same time, we have to keep a minimum of attributes so that the algorithms can perform efficiently. Also, we can analyze subgroups of attributes and tables, so that we can have segments.

Pre-Processing.- We have to decide whether we can keep ‘clean’ data or just handle the ‘noise’ in them.

Data-Mining.- Although very important, only consumes about 15%-20% of the time invested, compared to a very high percent of previous steps, specially planning. Algorithms used here, so, are very fast, and may be used several of them. Most of the tasks involved here and understand and select data, because the really useful data comes only after several tests.

Post-Processing.- Expert intervention in interpreting data.

To put it simply, KDD is an iterative process, which has complex interactions with heterogeneous tools, and in which success depends heavily on these as well on ‘the know-how’ of the dominion.

3. Discussion

For the number of applications discussed, there are some useful to the general public, others, at the businesses service. To me, it is important how to predict and prevent highway accidents; but is it useful to the public to be ‘hunted’ down by the marketing alchemists? After all, marketing does not always work on the public’s benefit. So, who is serving who? The problem with tracking cookies, for example, and privacy shall not be taken for granted. There is indeed a big sector of population who believes that some information shall remain secret and not used to hunt down clients. The concern has nothing to do with technology; is a concept more in the grounds of ethics.

I would like to see KDD in the benefit of mankind: to give a better service, to help people predict illnesses, to give information about where to find best traffic alternatives; not to gamble in Down Jones, not to be ‘targeted’ as a member of a group or client-prey.

4. Conclusions

As others have stated correctly in the past, we have reached a state where we can no longer take intelligent decisions without the help of the computers, for the amount of information has become so vast and complex, that in order to exploit its full capability we need to use machines to perform algorithms in the search for patterns.

It is desired, though, that human intervention may well be kept away of the data processing, and be let just for a machine to devour. But the main question should be not whether a machine must improve its intelligence to tackle this challenge, but for us to always do good use of the information we are given. Perhaps an humanistic view in the widely-technological culture we live in.