Chapter IV

The Discovery of Interesting Nuggets Using Heuristic Techniques

Beatriz de la Iglesia
Victor J. Rayward-Smith
University of East Anglia, UK

Knowledge Discovery in Databases (KDD) is an iterative and interactive process involving many steps (Debuse, de la Iglesia, Howard & Rayward-Smith, 2000). Data mining (DM) is defined as one of the steps in the KDD process. According to Fayyad, Piatetsky-Shapiro, Smyth and Uthurusamy (1996), there are various data mining tasks including: classification, clustering, regression, summarisation, dependency modeling, and change and deviation detection. However, there is a very important data mining problem identified previously by Riddle, Segal and Etzioni (1994) and very relevant in the context of commercial databases, which is not properly addressed by any of those tasks: nugget discovery. This task has also been identified as partial classification (Ali, Manganaris & Srikant, 1997). Nugget discovery can be defined as the search for relatively rare, but potentially important, patterns or anomalies relating to some pre-determined class or classes. Patterns of this type are called nuggets.

This chapter will present and justify the use of heuristic algorithms, namely Genetic Algorithms (GAs), Simulated Annealing (SA) and Tabu Search (TS), on the data mining task of nugget discovery. First, the concept of nugget discovery will be introduced. Then the concept of the interest of a nugget will be discussed. The necessary properties of an interest measure for nugget discovery will be presented. This will include a partial ordering of nuggets based on those properties. Some of the existing measures for nugget discovery will be reviewed in light of the properties established, and it will be shown that they do not display the required properties. A suitable evaluation function for nugget discovery, the fitness measure, will then be discussed and justified according to the required properties.
A number of algorithms, including the heuristic algorithms, will be introduced briefly. Experiments using those algorithms on some of the UCI repository databases (Merz & Murphy, 1998) will be reported. Conclusions about the suitability of the different algorithms on datasets with different characteristics can be drawn from these experiments. The three heuristics—Genetic Algorithms, Simulated Annealing and Tabu Search—will also be compared in terms of their implementation, results and performance.

**THE DATA MINING TASK OF NUGGET DISCOVERY**

In any KDD project, one of the first decisions that has to be made is what is the primary task that the user wants to achieve. The “high level” primary tasks of the KDD process are defined in the literature (Fayyad, Piatetsky-Shapiro & Smyth, 1996) as prediction and description. Prediction involves using some variables or fields in the database to predict unknown or future values of other variables of interest. Description focuses on finding human-interpretable patterns describing the data. The main distinction between prediction and description is who interprets the discovered knowledge. In prediction the system interprets the knowledge, whereas in description, it is the analyst or the user that interprets it.

Once the high-level goal of the process is established, the particular data mining task to be undertaken has to be chosen. This is known as the “low-level” task. As mentioned in the previous section, the most commonly recognised tasks are: classification, clustering, regression, summarisation, dependency modeling, and change and deviation detection. We will focus on the task of classification.

The type of data used for classification contains a pre-defined class assignment for each case or record in the database. This type of data is often encountered in commercial databases. The high-level goal of the user, when analysing this type of data, is sometimes prediction. This is when the user wants to infer a model that will allow him/her to assign a class to new data. For a predictive goal, a complete classification (that is, a complete model that assigns a class to each case or record in the database) may be necessary and appropriate. This would definitely fall under the heading of a classification task. An example of a classification task may be to build a decision tree (Quinlan, 1986) to differentiate between those customers that represent a good credit risk and those that do not, based on a database of financial information. The database must contain some classification of customers into good and bad credit risks, based on their past performance.

When the high-level goal is descriptive, it is not always necessary to provide a complete classification. This may indeed be detrimental to obtaining interesting and understandable patterns. The objective in many cases is to identify relatively rare, but potentially important, patterns or anomalies relating to some class or classes. We will call this type of pattern a nugget, and hence we will call this task nugget discovery. For instance, in the previous example, the bank may be
Preserving Privacy in Time Series Data Mining
Ye Zhu, Yongjian Fu and Huirong Fu (2011). *International Journal of Data Warehousing and Mining* (pp. 64-85).
www.igi-global.com/article/preserving-privacy-time-series-data/58638?camid=4v1a