Chapter 5 Summary

A data mining algorithm can be defined as a procedure that takes data and makes patterns or models in a finite number of steps. Defining a data mining algorithm in order to complete a particular task, we must also specify the task we are addressing, the model or pattern we are attempting to fit, the score function used to judge the quality of the fit, the method we are using to search the data, and the technique by which the data is stored. The components of deciding on the task, model or pattern, score function, are typically part of the human setup process whereas the remainder of the data mining algorithm is computational. The analysis of data mining algorithms describes existing data mining algorithms in an attempt to make it easier to compare different algorithms. The synthesis of data mining algorithms is the creation of data mining algorithms with different properties by combining existing components. The components of a data mining algorithm listed above provide a framework for both the analysis and the synthesis of data mining algorithms.

First consider an example of a data mining algorithm is the CART algorithm for building trees. It produces a tree-based structure of classification and regression models by employing a statistical procedure. For the components described above, the task of the CART algorithm is prediction or classification, the model is trees, the score function is the cross-validated loss function, the search method is the greedy local search, and the data management method is unspecified. The CART algorithm's defining feature is that of its model structure. The structure of the tree is not predefined but is derived from the data. The tree structure gives a hierarchical form which is what separates the CART algorithm from other classification algorithms without tree structures.

Looking back at the components of a data mining algorithm described above, we can think of the components as a "tuple" of *{model structure, score function, search method, data management technique}*. Looking at the components this way makes it easy to see that there is a large number of different data mining algorithms that can be generated by simply using these components. When we add in the idea that one data mining algorithm can combine different methods within each component, we see that there are potentially infinitely many different algorithms. We can bring this number down to a manageable amount by considering that there is only a small number of fundamental values for each component, meaning that for each component, there is only a few categories that are commonly used. This gives us a relatively

small number of different techniques by which a data mining algorithm is constructed. By looking at components of data mining algorithms, there is an emphasis placed on their core properties rather than the algorithm itself. When creating a data mining algorithm, we then look at the components which best fit our needs rather than trying to simply choose an existing algorithm.

Three well-known data mining algorithms are feedforward multilayer perceptrons or MLPs, the A Priori algorithm, and vector-space algorithms. MLPs are used to model nonlinear regression problems. Their task is to make a prediction and they are structured by nonlinear transformations of weighted sums taken from the inputs. MLPs have a score function of the sum of squared errors and use a search method of the steepest descent. In a MLP algorithm, the data management technique used is online or batch. The most defining feature of these algorithms is that they are multilayer and have a nonlinear model structure. The A Priori algorithm is used for association rule learning. An association rule is a probabilistic statement about occurrence of events with other events. The A Priori algorithm has a task of describing the associations between inputs and is structured by the probabilistic association rules. It uses thresholds on accuracy and support as its score function and takes a breadth-first approach to a systematic search method. In the A Priori algorithm, data is managed through multiple linear scans. The A Priori algorithm's defining feature is that of its search and data management components. Vector-space algorithms are generally used for text retrieval. More specifically, the task of a vector-space algorithm is the retrieval of the k documents in a database which are considered the most similar relative to the give query. The algorithm represents data as vectors and uses a score function of the angle between any two vectors. In vector-space algorithms various techniques are used as the search method along with the data management technique. Vector-space algorithms' defining feature is that of their representation of the data through the use of vectors.

Different data mining algorithms may have key differences between the techniques used, but they all consist of essentially the same components. Using these components as building blocks we are able to create an algorithm to achieve a given task with the goal of analyzing patterns and models within the data.