Chapter 4

The Scientific Data Mining Process

"When I *use a word," Humpty Dumpty said, in rather a scornful tone, "it means just what I choose it to mean—neither more nor less."*

—Lewis Carroll [87, p. 214]

In Chapter 2, I described various ways in which data mining was being applied to science and engineering problems. Next, in Chapter 3, I identified common themes across these diverse applications. In this chapter, I use these themes to describe an end-to-end data mining process, including various tasks at each step of the process and the order in which these tasks are usually performed to extract useful information from scientific data sets. These tasks will be discussed in more detail in later chapters. I will also comment on some overall characteristics of the data mining process and explain why I choose to define the process in a manner which differs from the more commonly used definition of data mining.

This chapter is organized as follows. Section 4.1 describes the end-to-end process of scientific data mining, motivated by the applications in Chapter 2 and the common themes identified in Chapter 3. Next, in Section 4.2, I make some general observations on the process, followed in Section 4.3 by my rationale for defining the data mining process to be broader than the common definition of the process. Section 4.4 concludes the chapter with a brief summary.

4.1 The tasks in the scientific data mining process

The description of scientific data types in Section 3.1 and the observations about the low-level nature of the raw scientific data discussed in Section 3.2.10 indicate that the raw data cannot be input directly to pattern recognition algorithms. This suggests that the data must first be preprocessed to bring it to a form suitable for pattern recognition. The common themes discussed in Chapter 3 suggest the tasks that might comprise this preprocessing of the data.

Figure 4.1 outlines one way in which these tasks might be incorporated into an end-to-end data mining system for analyzing data from a science or engineering application.

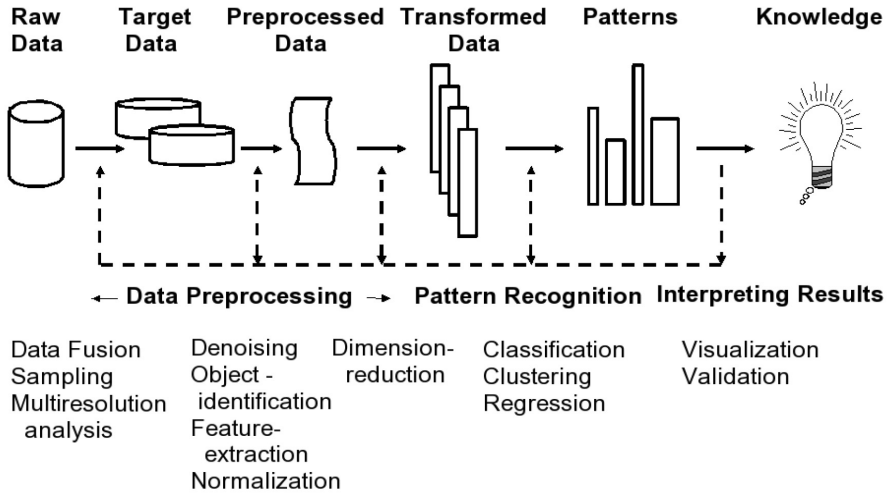

Figure 4.1. *The end-to-end scientific data mining process.*

Starting with the raw data in the form of images or meshes, we successively process these data into more refined forms, enabling further processing of the data and the extraction of relevant information. The terms, such as *Raw data* and *Target data*, used in Figure 4.1 are somewhat generic and also used in the mining of commercial data, where they are used to describe the process of *Knowledge Discovery in Databases* (KDD) [185]. In adapting this process to scientific data sets, I have retained the basic framework, but changed the tasks necessary to transform the data from one form to the next.

First, a few brief observations about the scientific data mining process; I will expand on these observations later, but it is helpful to keep them in mind while reading this section. The data mining process is essentially iterative and interactive—any one step can lead to the refinement of previous steps and the domain scientists should be actively involved in each step to validate the results. Not all of the tasks are needed in every problem and the order in which the steps are performed may change from one problem to another. The tasks listed are the ones I have found to be useful in many scientific applications; I am sure there are tasks I have missed as they may be specific to a particular problem. Each task can be addressed using several algorithms; these algorithms vary in their computational complexity, number of input parameters, suitability to a problem, and robustness to input parameters.

I next describe each of the steps in the scientific data mining process in more detail, followed by some general observations on the end-to-end process. I also discuss the ways in which the approach outlined in this chapter differs from mining of commercial data sets and the more traditional view of data mining as one step of the KDD process.

4.1.1 Transforming raw data into target data

The original or "raw" data which are provided for data mining often need extensive processing before they can be input to a pattern recognition algorithm. These algorithms typically require, as input, the objects (or data items) in the data set, with each object described by

a set of features. Thus, we first need to identify the objects in the data and extract features representing each object. In scientific data sets, the data may need to be processed before we can even identify the objects in the data. These processing steps may include tasks such as:

- **Data size reduction:** One task that is very helpful in the initial processing is to reduce the size of the data set. This is especially useful when the data set is very large and we are trying to understand the problem, doing exploratory analysis, or trying out different techniques for the various steps to determine a good solution approach. As we have observed in Chapter 2, scientific applications can result in very large data sets, and it is often helpful to first understand the data by working with a smaller sample and doing some exploratory analysis.

  One approach to reducing the size of the data is through sampling. During the initial analysis, we can select a small sample randomly from the full data set. Later on, as we begin formulating a solution approach, we need to increase the size of the data to reflect the variability which may be present naturally in the data. This would enable us to determine if the solution approach can handle this variability. Finally, most scientific problems require that a full pass be made through the entire data set. If the algorithms used in each task in the solution approach are robust enough, the final step can be fully automated.

  An alternative to sampling is to use multiresolution techniques, where we work with coarse resolution data. This may often be desirable in problems where the data are stored in a compressed form using an approach such as wavelets, which directly gives us a lower-resolution version of the data. Thus, the initial processing during the exploratory data analysis phase can be done on the lower-resolution data, with the full-resolution data being used in later iterations of the analysis.

  Of course, if the data set is massive, one can combine both the sampling and the multiresolution approaches. These techniques for reducing the size of the data are discussed further in Chapter 5.

- **Data fusion:** If the scientific data have been obtained from different sensors, they must be fused before the complementary information in the different data can be exploited. In the case of image data, the step of data fusion may include registration, which is the process of aligning two images of a scene or object, taken at different times or at different resolutions. For example, one frame of a video of a scene may need to be registered with the frame taken the very next instant. This is because even a stationary camera may move a little, resulting in successive frames of the video being misaligned by a pixel or two. Or, satellite images, obtained to understand how the vegetation in a region has changed over the years, must be registered so scientists can map corresponding areas for change detection.

  A different type of data fusion is required when the data are of different modalities, for example, text and images, or video and audio. In this case, fusion may occur at the feature level, where the different modalities of data each contribute features describing an object. For example, a technical document may be represented by text features representing the words in the document, image features representing the images in the document, and other features representing the tables in the document. We can also have fusion at the decision level, where each modality is processed independently, and the final decision made by appropriately combining the decisions

from the different modalities. In both these cases, the features or decision may be appropriately weighted to account for the quality of data in each modality.

These issues of data fusion, including data registration, are discussed in more detail in Chapter 6.

- **Image enhancement:** As we have observed, the presence of noise can make it difficult to analyze the data. This is especially true in the case of images if the objects of interest are occluded or not clearly separated from the background due to low contrast; in such cases, it may be difficult to extract the objects from the image.

  Though there are general techniques to reduce noise in images or improve their contrast, many of the techniques used are domain specific. This is because the noise in the data may be due to domain effects, such as the sensor characteristics, or due to external circumstances, such as atmospheric turbulence in astronomical data sets. These noise characteristics may differ across images in a data set, making it difficult to use fully automated techniques to remove the noise. Also, any processing to reduce noise from an image can adversely affect the signal, and appropriate trade-offs may be necessary to reduce the noise or improve the contrast, while minimizing the effect on the signal.

  Techniques for enhancing images are discussed further in Chapter 7.

4.1.2 Transforming target data into preprocessed data

Once the data have been reduced through sampling and/or multiresolution techniques, complementary data sources have been fused, and the data enhanced to make it easier to extract the objects of interest, potential next steps in the scientific data mining process include:

- **Object identification:** Here, the objects of interest are identified in the data, for example, by isolating the pixels which form a galaxy from the background. This step is often time consuming for images and meshes as many different techniques may need to be tried. In addition, depending on the quality of the data, the parameters used in the algorithms may need to be changed, making completely automated analysis difficult.

  In some problems, the identification of objects may be relatively easy, for example, in the case of a physics experiment where we are interested in classifying certain types of events, an "object" may be the time window in which each event occurs. In other problems, the objects may be poorly defined to the extent that defining the objects of interest is one of the main goals of the data mining endeavor. For example, if the physical phenomenon is poorly understood, the scientists may be unsure of what they are looking for in the data. In other problems, the object of interest may evolve over time, splitting and merging with other objects, making it difficult to have a single definition of what constitutes an "object." Or, the data structure used to store the data may preclude easy identification of the object, for example, in an unstructured mesh. In addition, if the data have been generated using a parallel computer, they may be distributed over several files, making it difficult to identify an object split across many files.

Techniques for object identification, especially for images and meshes, are discussed further in Chapter 8.

- **Feature extraction:** As observed earlier, the word "feature" is used to imply different things in different domains. In this book, I use "feature" to mean any low-level attribute or measurement which can be extracted from the data. Features are typically used to represent the objects in the data and are extracted after the objects have been identified. These features are then used in the pattern recognition step.

  The features extracted for an object are usually scale-, rotation-, and translation-invariant. This is because the patterns in the data usually remain unchanged when they are scaled, rotated, or translated. For a feature to be useful, it must be robust, that is, it must be insensitive to small changes in the data. Often, we may need to try different ways of extracting the same feature as some techniques may yield features which are more consistent across the objects of interest.

  In addition to features representing the objects themselves, we frequently need to include "housekeeping" features for each object. These represent the metadata and include information such as the image in which the object was found, the location of the object in the image, the resolution of the image, and so on.

  The features which are extracted for an object are very dependent on the patterns of interest. For example, if the pattern reflects the shape of an object, then various shape-based features should be extracted. Often, it helps to extract features reflecting different properties of an object such as shape, texture, and statistical distributions of intensities, as it may not be obvious at first which type of feature is likely to be most discriminating for the pattern recognition task at hand.

  Different types of features are discussed further in Chapter 9.

- **Normalization and cleanup of features:** Once we have identified the objects in the data, and extracted features representing the objects, as well as features representing the metadata for the objects, then we have a data item (or object) by feature matrix representing the data set. Each data item in this matrix is described by its feature vector. It is often helpful to process these features to ensure they are consistent and accurately represent the data. For example, the features may need to be normalized as the units corresponding to some features may make them appear more important than other features. This normalization may need to take into account the number of features corresponding to each feature type, for example, the number of Gabor texture features may be much larger than the number of features representing the histogram of the intensity of the pixels in the object.

  In addition to normalization, we should check that the features representing an object are valid and complete. For example, features may be missing because a sensor was inoperational. Or, feature values may be outliers and may not have a physical meaning due to inaccuracies in the collection or the input of the data values. Such features must be appropriately processed else they may adversely affect the step of pattern recognition. Techniques for doing this are discussed further in Chapter 9.

4.1.3 Converting preprocessed data into transformed data

Once the initial steps of preprocessing have been applied to the raw data, the objects in the data identified, and features representing them extracted, we have a matrix where the rows represent the data items or objects and the columns represent the features. We could also have considered the rows to be the features and the columns to be the objects, but in this book, we will focus on the data item by feature matrix. In the initial stages of the analysis, we may not consider all the objects, but only those found in the subset of data being analyzed. In addition, we may often extract far more features than necessary to represent each object, as we may not know which features are the most discriminating. These features may number in the tens or even the hundreds or thousands.

Reducing the number of features representing each object is an important step prior to pattern recognition. There are several reasons why this can be helpful:

- Some of the pattern recognition tasks (such as finding nearest neighbors or classification) can be considered to be a search through a high-dimensional space, where the dimension is the number of features. For example, in classification, we need to find the hyperplanes which separate, say the positive examples from the negative ones. If each example, or data item, is represented by many features, trying to determine the appropriate hyperplanes can be difficult. The *curse of dimensionality*, as observed by Bellman [31], indicates that in the absence of simplifying assumptions, the sample size needed to estimate a function of several variables to a given degree of accuracy grows exponentially with the number of variables. In scientific data, where training sets are usually generated manually, large sample sizes may not be an option in classification problems.

- Some of the features extracted to represent an object may be irrelevant (such as the location in the file from which the object was extracted), while others may not be discriminating enough (such as the color feature in a sample of all red objects). While the former may be necessary for identifying where an object came from, we could save on computational time by not extracting features unless they are discriminating enough. However, this assumes that the initial sample of data is representative enough so that a feature, once discarded, does not turn out to be discriminating when the entire data set is considered.

- Reducing the number of features may improve the ability to visualize or interpret the patterns in the data. If the number of features is reduced to a small number, say 3–10, it may be possible to use techniques such as three-dimensional visualization or parallel plots (see Chapter 12) to understand how the data items are grouped together based on the features.

- By reducing the number of features, we can reduce both the computational time and the memory requirements of the pattern recognition algorithms which are applied subsequent to the dimension reduction.

- In some scientific domains, the features extracted from the images or meshes may be stored in databases. Efficient indexing schemes for retrieval of these data are feasible when the number of dimensions is 8–12 (see [209, 177]). Dimension reduction would therefore provide more efficient access to the feature data for analysis.

There are several ways in which the number of features representing an object can be reduced. These include both techniques which transform the data into a lower-dimensional space as well as techniques which select a subset of the features. I will discuss these techniques in more detail in Chapter 10.

4.1.4 Converting transformed data into patterns

Once the data items are represented by a possibly reduced set of features, we can use this transformed data to identify patterns. Depending on the problem, this identification of patterns can take several forms, including:

- **Classification and regression:** If we have a class label associated with each data item, that is, we have a training set, we can use it to build a model of the data which separates one class label from another. There are several classification algorithms, each of which builds a different type of model, and each with its pros and cons. If the label is continuous instead of discrete, the problem can be addressed using regression techniques.
- **Clustering:** If there are no class labels associated with the objects, then we can use clustering techniques to group the objects based on their similarity in feature space. The objects in each cluster are then analyzed to determine what characteristics brought them together.
- **Information retrieval:** A related task is to retrieve objects which are similar to a query object. This can be done by identifying objects which are close to the query in feature space. If the features are an accurate representation of the objects, then objects which are close to the query in feature space are likely to be similar to it.
- **Outlier or anomaly detection:** In some problems, we may be interested in identifying the data items which are anomalous; that is, they do not belong with the rest of the items. Such problems typically arise in streaming data, where near-real-time analysis is required.
- **Association rules:** These techniques are popular in commercial data in the context of market basket analysis where they are used to determine the types of items which are often bought together. Association rules are just making inroads into scientific data mining in domains such as bioinformatics.
- **Tracking:** In spatiotemporal data, the features associated with an object, such as its size and location, can be used to track the object from one time step to the next.

Scientific problems often have characteristics which can make the direct application of various pattern recognition techniques a challenge. I will discuss these in more detail in Chapter 11.

4.1.5 Converting patterns into knowledge

Once the patterns in the data have been identified, they must be visualized and presented to the scientist for validation. This may result in the iterative refinement of one or more of the steps in the data mining process.

This visual display of the results has to be done with care to maintain any relationships between the objects. This can be difficult when the objects are in a high-dimensional space and they have to be projected into two or three dimensions for display. Another problem arises when the size of the data set is so large that a typical display is insufficient.

This visual display of information is not restricted to just the final step in the data mining process. In fact, it is often useful to apply the same techniques for information visualization during the initial exploratory analysis of the data and in evaluating the results from the intermediate steps. Such validation can often yield important insights into the data and suggest appropriate solution approaches for the next steps. I will elaborate on the visualization and validation of results in Chapter 12.

4.2 General observations about the scientific data mining process

In Section 4.1, I briefly described the various tasks which comprise the overall process of scientific data mining. Next, I make some general observations about the process of mining science and engineering data.

- The data flow diagram presented in Figure 4.1 is one that I have found to cover the needs of many of the scientific applications I have encountered. It is by no means the only approach to the analysis of scientific data sets; variations and enhancements may be necessary as required by an application. For example, the order in which the tasks are done may change from one application to another. Depending on the quality of the data, it may be necessary to enhance image data from different sensors before fusing them as the enhancement may make the fusion algorithms less sensitive to the setting of the parameters. Also, not all tasks may be necessary in all applications; for example, denoising data is not a key part of the analysis of data from computer simulations, but is often required for experimental data. Tasks in addition to the ones described in Section 4.1 may be required as appropriate for specific problems, for example, to incorporate domain-specific information.

- The data size shrinks as we move to the right from the raw data to the patterns identified in the data. The raw data may be in the form of images, while the pattern may be a simple model in the form of a decision tree, which can be used to label the objects in the images. If the original data are being processed using a parallel computer, then the shrinking of the data size would imply that it would be inefficient to use the same number of processors throughout the implementation of the entire process. In addition, there could also be issues of load balancing. For example, if we are analyzing image data, each processor could work on a subset of images. The denoising and enhancement of the images, as well as the identification of the objects in the images, could be done in parallel. However, if some images had far more objects than others, the processors working on those images would have more work to do in the feature extraction step than the processors working on images with few objects. Further, if more than one processor is processing a single image, some processors may have no objects in their subimage, while other processors may have many. An object could also be split between the subimages assigned to different processors, requiring appropriate merging of the processed data.

- The scientific data mining process is very interactive. The domain scientists are involved in every step, starting with the initial, perhaps tentative, formulation of the problem to providing information on the noise characteristics of the sensors for use in enhancing the data, validating the objects identified in the data, identifying robust ways of calculating the features for the objects, selecting features which may or may not be relevant to the problem, and most importantly, validating the results obtained at each step.

  The scientific data mining process is also an iterative process. The results of any one step may indicate that a previous step needs to be refined. For example, identification of the objects in image data may indicate that the noise in the images has to be reduced further in order to separate an object from the background. Or, the step of pattern recognition may indicate that some features which are key to discrimination are not rotation invariant and therefore objects which are rotated versions of objects in the training set are not being labeled correctly. Or, the error rate of the pattern recognition step could be high, indicating that the features extracted are not representative enough of the patterns being considered or that the quality of the training data could be improved.

  If the data set is very large, a practical approach to analyzing it might require working initially with a small subset of the data, processing these data using the initial steps of Figure 4.1, and when a set of algorithms is found to work, processing additional data to see if the quality of the results still hold and if it is possible to proceed further along the analysis pipeline. If the initial sample is not representative enough, the new data added may indicate a different choice of algorithms or parameter setting for the algorithms.

- The tasks in Figure 4.1 can be implemented using several algorithms. For example, there are several ways of reducing noise in image data and several different classification algorithms to create a model given the training data. These algorithms differ in their suitability to a problem, the assumptions they make about the data, their computational complexity, the accuracy of the results, their robustness to input parameters, and their interpretability. Often, several algorithms may need to be tried before one suitable for the data and problem is found. In some cases, we may even need to design algorithms which are tuned to the characteristics of the data or problem.

- Each task in the process described in Figure 4.1 generates some output which can be directly input to the next step. However, it is often advantageous to save the results which are output at each step. It allows us to experiment with different algorithms for a step without having to repeat all the previous steps. Also, if new data become available, they can easily be incorporated into the process by repeating only the steps which are absolutely necessary, without processing all the data twice. Saving intermediate data essentially makes it easier to handle the iterative nature of the scientific data mining process.

4.3 Defining scientific data mining: The rationale

As I discussed in Chapter 1, there are several definitions of data mining, with some considering it to be one of the steps in the Knowledge Discovery in Databases (KDD) process in which patterns are extracted from data, and others considering it to be the overall process of

extracting useful information from data. When this data arises from science and engineering applications, it is clear that we cannot analyze the data by directly applying algorithms which extract patterns from this data.

As we have seen in Chapter 2, the data in many scientific applications are in the form of images or mesh data from simulations, where the values of the variables are available at each pixel or mesh point. However, we are interested in patterns among the objects in the data, objects which are at a higher level, for example a galaxy in an astronomical image, where the galaxy is a collection of pixels. Going from the data at the level of pixels, or mesh points, to data which are suitable for pattern recognition, is a very time-consuming process, involving trial and error. As I have described in this chapter, there are several tasks which must be performed before the data are ready for pattern recognition. In many problems, these tasks typically take 80–90% of the total time spent in analyzing the data. Often, the quality of the final results is very dependent on the quality of the initial preprocessing. Further, these tasks provide the means of obtaining input from the domain experts, input which is key not only to understanding the problem, but also solving it. For these reasons, I consider data mining to be the overall process of extracting information from data, not just one step in the KDD process. Of course, one might also say that for many scientific applications, the data are not stored in databases, and the term KDD might not be applicable in the strictest sense of the term.

4.4 Summary

In this chapter, I have described the tasks involved in a typical end-to-end process for the analysis of scientific data sets, where one starts with the raw data in the form of images or meshes and extracts useful information. The process may involve all or some of the following tasks: sampling, multiresolution, data fusion, object identification, feature extraction, dimension reduction, pattern recognition, visualization, and validation. I shall describe how we can accomplish these tasks in the next several chapters. I start by describing ways in which we can reduce the size of the data to make the initial exploratory analysis feasible for massive data sets.