CHAPTER 1

THE IMPORTANCE OF FEATURE SELECTION

1.1 KNOWLEDGE DISCOVERY

It is estimated that every 20 months or so the amount of information in the world doubles. In the same way, tools for use in the various knowledge fields (acquisition, storage, retrieval, maintenance, etc.) must develop to combat this growth. Knowledge is only valuable when it can be used efficiently and effectively; therefore knowledge management is increasingly being recognized as a key element in extracting its value. This is true both within the research, development, and application of computational intelligence and beyond.

Central to this issue is the knowledge discovery process, particularly knowledge discovery in databases (KDD) [10,90,97,314]. KDD is the nontrivial process of identifying valid, novel, potentially useful, and ultimately understandable patterns in data. Traditionally data was turned into knowledge by means of manual analysis and interpretation. For many applications manual probing of data is slow, costly, and highly subjective. Indeed, as data volumes grow dramatically, manual data analysis is becoming completely impractical in many domains. This motivates the need for efficient, automated knowledge discovery. The KDD process can be decomposed into the following steps, as illustrated in Figure 1.1:

- *Data selection.* A target dataset is selected or created. Several existing datasets may be joined together to obtain an appropriate example set.

Computational Intelligence and Feature Selection: Rough and Fuzzy Approaches, by Richard Jensen and Qiang Shen

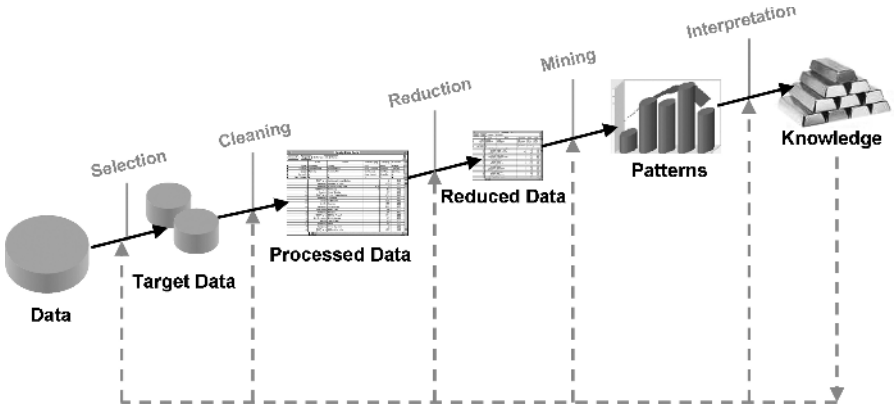

Figure 1.1 Knowledge discovery process (adapted from [97])

- *Data cleaning/preprocessing.* This phase includes, among other tasks, noise removal/reduction, missing value imputation, and attribute discretization. The goal is to improve the overall quality of any information that may be discovered.
- *Data reduction.* Most datasets will contain a certain amount of redundancy that will not aid knowledge discovery and may in fact mislead the process. The aim of this step is to find useful features to represent the data and remove nonrelevant features. Time is also saved during the data-mining step as a result.
- *Data mining.* A data-mining method (the extraction of hidden predictive information from large databases) is selected depending on the goals of the knowledge discovery task. The choice of algorithm used may be dependent on many factors, including the source of the dataset and the values it contains.
- *Interpretation/evaluation.* Once knowledge has been discovered, it is evaluated with respect to validity, usefulness, novelty, and simplicity. This may require repeating some of the previous steps.

The third step in the knowledge discovery process, namely data reduction, is often a source of significant data loss. It is this step that forms the focus of attention of this book. The high dimensionality of databases can be reduced using suitable techniques, depending on the requirements of the future KDD processes. These techniques fall into one of two categories: those that transform the underlying meaning of the data features and those that preserve the semantics. Feature selection (FS) methods belong to the latter category, where a smaller set of the original features is chosen based on a subset evaluation function. In knowledge discovery, feature selection methods are particularly desirable as these facilitate the interpretability of the resulting knowledge.

1.2 FEATURE SELECTION

There are often many features in KDD, and combinatorially large numbers of feature combinations, to select from. Note that the number of feature subset combinations with m features from a collection of N total features can be extremely large (with this number being $N!/[m!(N-m)!]$ mathematically). It might be expected that the inclusion of an increasing number of features would increase the likelihood of including enough information to distinguish between classes. Unfortunately, this is not true if the size of the training dataset does not also increase rapidly with each additional feature included. This is the so-called curse of dimensionality [26]. A high-dimensional dataset increases the chances that a data-mining algorithm will find spurious patterns that are not valid in general. Most techniques employ some degree of reduction in order to cope with large amounts of data, so an efficient and effective reduction method is required.

1.2.1 The Task

The task of feature selection is to select a subset of the original features present in a given dataset that provides most of the useful information. Hence, after selection has taken place, the dataset should still have most of the important information still present. In fact, good FS techniques should be able to detect and ignore noisy and misleading features. The result of this is that the dataset quality might even *increase* after selection.

There are two feature qualities that must be considered by FS methods: relevancy and redundancy. A feature is said to be relevant if it is predictive of the decision feature(s); otherwise, it is irrelevant. A feature is considered to be redundant if it is highly correlated with other features. An informative feature is one that is highly correlated with the decision concept(s) but is highly uncorrelated with other features (although low correlation does not mean absence of relationship). Similarly subsets of features should exhibit these properties of relevancy and nonredundancy if they are to be useful.

In [171] two notions of feature relevance, strong and weak relevance, were defined. If a feature is strongly relevant, this implies that it cannot be removed from the dataset without resulting in a loss of predictive accuracy. If it is weakly relevant, then the feature may sometimes contribute to accuracy, though this depends on which other features are considered. These definitions are independent of the specific learning algorithm used. However, this no guarantee that a relevant feature will be useful to such an algorithm.

It is quite possible for two features to be useless individually, and yet highly predictive if taken together. In FS terminology, they may be both redundant and irrelevant on their own, but their combination provides invaluable information. For example, in the exclusive-or problem, where the classes are not linearly separable, the two features on their own provide no information concerning this separability. It is also the case that they are uncorrelated with each other. However, when taken together, the two features are highly informative and can provide

good class separation. Hence in FS the search is typically for high-quality feature subsets, and not merely a ranking of features.

1.2.2 The Benefits

There are several potential benefits of feature selection:

1. *Facilitating data visualization*. By reducing data to fewer dimensions, trends within the data can be more readily recognized. This can be very important where only a few features have an influence on data outcomes. Learning algorithms by themselves may not be able to distinguish these factors from the rest of the feature set, leading to the generation of overly complicated models. The interpretation of such models then becomes an unnecessarily tedious task.
2. *Reducing the measurement and storage requirements*. In domains where features correspond to particular measurements (e.g., in a water treatment plant [322]), fewer features are highly desirable due to the expense and time-costliness of taking these measurements. For domains where large datasets are encountered and manipulated (e.g., text categorization [162]), a reduction in data size is required to enable storage where space is an issue.
3. *Reducing training and utilization times*. With smaller datasets, the runtimes of learning algorithms can be significantly improved, both for training and classification phases. It can sometimes be the case that the computational complexity of learning algorithms even prohibits their application to large problems. This is remedied through FS, which can reduce the problem to a more manageable size.
4. *Improving prediction performance*. Classifier accuracy can be increased as a result of feature selection, through the removal of noisy or misleading features. Algorithms trained on a full set of features must be able to discern and ignore these attributes if they are to produce useful, accurate predictions for unseen data.

For those methods that extract knowledge from data (e.g., rule induction) the benefits of FS also include improving the readability of the discovered knowledge. When induction algorithms are applied to reduced data, the resulting rules are more compact. A good feature selection step will remove unnecessary attributes which can affect both rule comprehension and rule prediction performance.

1.3 ROUGH SETS

The use of rough set theory (RST) [261] to achieve data reduction is one approach that has proved successful. Over the past 20 years RST has become a topic of great interest to researchers and has been applied to many domains (e.g.,

classification [54,84,164], systems monitoring [322], clustering [131], and expert systems [354]; see *LNCS Transactions on Rough Sets* for more examples). This success is due in part to the following aspects of the theory:

- Only the facts hidden in data are analyzed.
- No additional information about the data is required such as thresholds or expert knowledge on a particular domain.
- It finds a minimal knowledge representation.

The work on RST offers an alternative, and formal, methodology that can be employed to reduce the dimensionality of datasets, as a preprocessing step to assist any chosen modeling method for learning from data. It helps select the most information-rich features in a dataset, without transforming the data, all the while attempting to minimize information loss during the selection process. Computationally, the approach is highly efficient, relying on simple set operations, which makes it suitable as a preprocessor for techniques that are much more complex. Unlike statistical correlation-reducing approaches [77], it requires no human input or intervention. Most importantly, it also retains the semantics of the data, which makes the resulting models more transparent to human scrutiny.

Combined with an automated intelligent modeler, say a fuzzy system or a neural network, the feature selection approach based on RST not only can retain the descriptive power of the learned models but also allow simpler system structures to reach the knowledge engineer and field operator. This helps enhance the interoperability and understandability of the resultant models and their reasoning.

As RST handles only one type of imperfection found in data, it is complementary to other concepts for the purpose, such as fuzzy set theory. The two fields may be considered analogous in the sense that both can tolerate inconsistency and uncertainty—the difference being the type of uncertainty and their approach to it. Fuzzy sets are concerned with vagueness; rough sets are concerned with indiscernibility. Many deep relationships have been established, and more so, most recent studies have concluded at this complementary nature of the two methodologies, especially in the context of granular computing. Therefore it is desirable to extend and hybridize the underlying concepts to deal with additional aspects of data imperfection. Such developments offer a high degree of flexibility and provide robust solutions and advanced tools for data analysis.

1.4 APPLICATIONS

As many systems in a variety of fields deal with datasets of large dimensionality, feature selection has found wide applicability. Some of the main areas of application are shown in Figure 1.2.

Feature selection algorithms are often applied to optimize the classification performance of image recognition systems [158,332]. This is motivated by a peaking phenomenon commonly observed when classifiers are trained with a limited

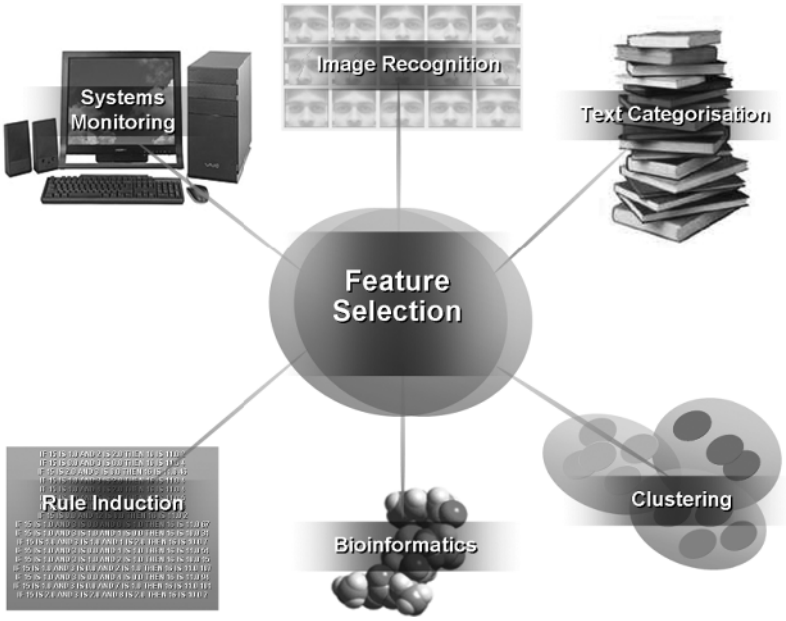

Figure 1.2 Typical feature selection application areas

set of training samples. If the number of features is increased, the classification rate of the classifier decreases after a peak. In melanoma diagnosis, for instance, the clinical accuracy of dermatologists in identifying malignant melanomas is only between 65% and 85% [124]. With the application of FS algorithms, automated skin tumor recognition systems can produce classification accuracies above 95%.

Structural and functional data from analysis of the human genome have increased many fold in recent years, presenting enormous opportunities and challenges for AI tasks. In particular, gene expression microarrays are a rapidly maturing technology that provide the opportunity to analyze the expression levels of thousands or tens of thousands of genes in a single experiment. A typical classification task is to distinguish between healthy and cancer patients based on their gene expression profile. Feature selectors are used to drastically reduce the size of these datasets, which would otherwise have been unsuitable for further processing [318,390,391]. Other applications within bioinformatics include QSAR [46], where the goal is to form hypotheses relating chemical features of molecules to their molecular activity, and splice site prediction [299], where junctions between coding and noncoding regions of DNA are detected.

The most common approach to developing expressive and human readable representations of knowledge is the use of *if-then* production rules. Yet real-life problem domains usually lack generic and systematic expert rules for mapping feature patterns onto their underlying classes. In order to speed up the rule

induction process and reduce rule complexity, a selection step is required. This reduces the dimensionality of potentially very large feature sets while minimizing the loss of information needed for rule induction. It has an advantageous side effect in that it removes redundancy from the historical data. This also helps simplify the design and implementation of the actual pattern classifier itself, by determining what features should be made available to the system. In addition the reduced input dimensionality increases the processing speed of the classifier, leading to better response times [12,51].

Many inferential measurement systems are developed using data-based methodologies; the models used to infer the value of target features are developed with real-time plant data. This implies that inferential systems are heavily influenced by the quality of the data used to develop their internal models. Complex application problems, such as reliable monitoring and diagnosis of industrial plants, are likely to present large numbers of features, many of which will be redundant for the task at hand. Additionally there is an associated cost with the measurement of these features. In these situations it is very useful to have an intelligent system capable of selecting the most relevant features needed to build an accurate and reliable model for the process [170,284,322].

The task of text clustering is to group similar documents together, represented as a bag of words. This representation raises one severe problem: the high dimensionality of the feature space and the inherent data sparsity. This can significantly affect the performance of clustering algorithms, so it is highly desirable to reduce this feature space size. Dimensionality reduction techniques have been successfully applied to this area—both those that destroy data semantics and those that preserve them (feature selectors) [68,197].

Similar to clustering, text categorization views documents as a collection of words. Documents are examined, with their constituent keywords extracted and rated according to criteria such as their frequency of occurrence. As the number of keywords extracted is usually in the order of tens of thousands, dimensionality reduction must be performed. This can take the form of simplistic filtering methods such as word stemming or the use of stop-word lists. However, filtering methods do not provide enough reduction for use in automated categorizers, so a further feature selection process must take place. Recent applications of FS in this area include Web page and bookmark categorization [102,162].

1.5 STRUCTURE

The rest of this book is structured as follows (see Figure 1.3):

- *Chapter 2: Set Theory.* A brief introduction to the various set theories is presented in this chapter. Essential concepts from classical set theory, fuzzy set theory, rough set theory, and hybrid fuzzy-rough set theory are presented and illustrated where necessary.

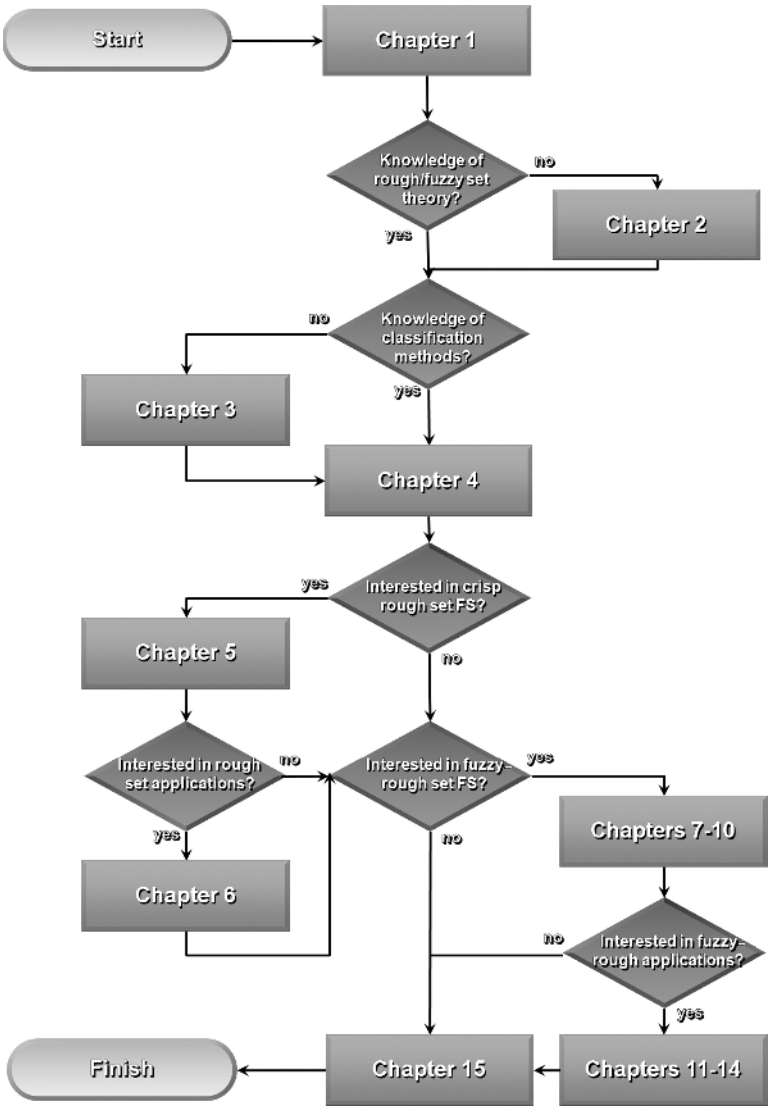

Figure 1.3 How to read this book

- *Chapter 3: Classification Methods.* This chapter discusses both crisp and fuzzy methods for the task of classification. Many of the methods presented here are used in systems later in the book.
- *Chapter 4: Dimensionality Reduction.* A systematic overview of current techniques for dimensionality reduction with a particular emphasis on feature selection is given in this chapter. It begins with a discussion of those

reduction methods that irreversibly transform data semantics. This is followed by a more detailed description and evaluation of the leading feature selectors presented in a unified algorithmic framework. A simple example illustrates their operation.

- *Chapter 5: Rough Set-based Approaches to Feature Selection.* This chapter presents an overview of the existing research regarding the application of rough set theory to feature selection. Rough set attribute reduction (RSAR), the precursor to the developments in this book, is described in detail. However, these methods are unsuited to the problems discussed in Section 5.11. In particular, they are unable to handle noisy or real-valued data effectively—a significant problem if they are to be employed within real-world applications.
- *Chapter 6: Applications I: Use of RSAR.* This chapter looks at the applications of RSAR in several challenging domains: medical image classification, text categorization, and algae population estimation. Details of each classification system are given with several comparative studies carried out that investigate RSAR's utility. Additionally a brief introduction to other applications that use a crisp rough set approach is provided for the interested reader.
- *Chapter 7: Rough and Fuzzy Hybridization.* There has been great interest in developing methodologies that are capable of dealing with imprecision and uncertainty. The large amount of research currently being carried out in fuzzy and rough sets is representative of this. Many deep relationships have been established, and recent studies have concluded at the complementary nature of the two methodologies. Therefore it is desirable to extend and hybridize the underlying concepts to deal with additional aspects of data imperfection. Such developments offer a high degree of flexibility and provide robust solutions and advanced tools for data analysis. A general survey of this research is presented in the chapter, with a focus on applications of the theory to disparate domains.
- *Chapter 8: Fuzzy-Rough Feature Selection.* In this chapter the theoretical developments behind this new feature selection method are presented together with a proof of generalization. This novel approach uses fuzzy-rough sets to handle many of the problems facing feature selectors outlined previously. A complexity analysis of the main selection algorithm is given. The operation of the approach and its benefits are shown through the use of two simple examples. To evaluate this new fuzzy-rough measure of feature significance, comparative investigations are carried out with the current leading significance measures.
- *Chapter 9: New Developments of FRFS.* Fuzzy-rough set-based feature selection has been shown to be highly useful at reducing data dimensionality, but possesses several problems that render it ineffective for datasets possessing tens of thousands of features. This chapter presents three new approaches to fuzzy-rough feature selection (FRFS) based on

fuzzy similarity relations. The first employs the new similarity-based fuzzy lower approximation to locate subsets. The second uses boundary region information to guide search. Finally, a fuzzy extension to crisp discernibility matrices is given in order to discover fuzzy-rough subsets. The methods are evaluated and compared using benchmark data.

- *Chapter 10: Further Advanced FS Methods.* This chapter introduces two promising areas in feature selection. The first, feature grouping, is developed from recent work in the literature where groups of features are selected simultaneously. By reasoning with fuzzy labels, the search process can be made more intelligent allowing various search strategies to be employed. The second, ant-based feature selection, seeks to address the nontrivial issue of finding the smallest optimal feature subsets. This approach to feature selection uses artificial ants and pheromone trails in the search for the best subsets. Both of these developments can be applied within feature selection, in general, but are applied to the specific problem of subset search within FRFS in this book.
- *Chapter 11: Applications II: Web Content Categorization.* With the explosive growth of information on the Web, there is an abundance of information that must be dealt with effectively and efficiently. This area, in particular, deserves the attention of feature selection due to the increasing demand for high-performance intelligent Internet applications. This motivates the application of FRFS to the automatic categorization of user bookmarks/favorites and Web pages. The results show that FRFS significantly reduces data dimensionality by several orders of magnitude with little resulting loss in classification accuracy.
- *Chapter 12: Applications III: Complex Systems Monitoring.* Complex application problems, such as reliable monitoring and diagnosis of industrial plants, are likely to present large numbers of features, many of which will be redundant for the task at hand. With the use of FRFS, these extraneous features can be removed. This not only makes resultant rulesets generated from such data much more concise and readable but can reduce the expense due to the monitoring of redundant features. The monitoring system is applied to water treatment plant data, producing better classification accuracies than those resulting from the full feature set and several other reduction methods.
- *Chapter 13: Applications IV: Algae Population Estimation.* Biologists need to identify and isolate the chemical parameters of rapid algae population fluctuations in order to limit their detrimental effect on the environment. This chapter describes an estimator of algae populations, a hybrid system involving FRFS that approximates, given certain water characteristics, the size of algae populations. The system significantly reduces computer time and space requirements through the use of feature selection. The results show that estimators using a fuzzy-rough feature selection step produce more accurate predictions of algae populations in general.
- *Chapter 14: Applications V: Forensic Glass Analysis.* The evaluation of glass evidence in forensic science is an important issue. Traditionally this

has depended on the comparison of the physical and chemical attributes of an unknown fragment with a control fragment. A high degree of discrimination between glass fragments is now achievable due to advances in analytical capabilities. A random effects model using two levels of hierarchical nesting is applied to the calculation of a likelihood ratio (LR) as a solution to the problem of comparison between two sets of replicated continuous observations where it is unknown whether the sets of measurements shared a common origin. This chapter presents the investigation into the use of feature evaluation for the purpose of selecting a single variable to model without the need for expert knowledge. Results are recorded for several selectors using normal, exponential, adaptive, and biweight kernel estimation techniques. Misclassification rates for the LR estimators are used to measure performance.

- *Chapter 15: Supplementary Developments and Investigations.* This chapter offers initial investigations and ideas for further work, which were developed concurrently with the ideas presented in the previous chapters. First, the utility of using the problem formulation and solution techniques from propositional satisfiability for finding rough set reducts is considered. This is presented with an initial experimental evaluation of such an approach, comparing the results with a standard rough set-based algorithm, RSAR. Second, the possibility of universal reducts is proposed as a way of generating more useful feature subsets. Third, fuzzy decision tree induction based on the fuzzy-rough metric developed in this book is proposed. Other proposed areas of interest include fuzzy-rough clustering and fuzzy-rough fuzzification optimization.