Data Reduction with Rough Sets

Richard Jensen
Aberystwyth University, UK

Qiang Shen
Aberystwyth University, UK

INTRODUCTION

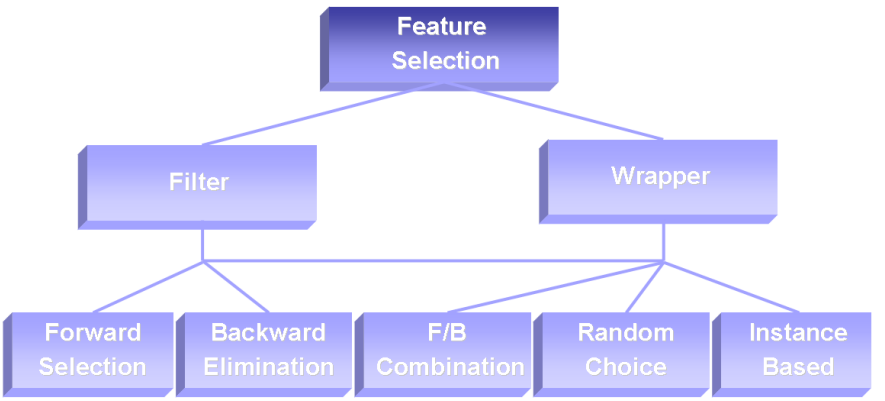
Data reduction is an important step in knowledge discovery from data. The high dimensionality of databases can be reduced using suitable techniques, depending on the requirements of the data mining processes. These techniques fall in to one of the following categories: those that transform the underlying meaning of the data features and those that are semantics-preserving. 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. The process aims to determine a minimal feature subset from a problem domain while retaining a suitably high accuracy in representing the original features. In knowledge discovery, feature selection methods are particularly desirable as they facilitate the interpretability of the resulting knowledge. For this, rough set theory has been successfully used as a tool that enables the discovery of data dependencies and the reduction of the number of features contained in a dataset using the data alone, while requiring no additional information.

BACKGROUND

The main aim of feature selection is to determine a minimal feature subset from a problem domain while retaining a suitably high accuracy in representing the original features. In many real world problems FS is a must due to the abundance of noisy, irrelevant or misleading features. A detailed review of feature selection techniques devised for classification tasks can be found in (Dash & Liu, 1997).

The usefulness of a feature or feature subset is determined by both its relevancy and its redundancy. A feature is said to be relevant if it is predictive of the decision feature(s) (i.e. dependent variable(s)), otherwise it is irrelevant. A feature is considered to be redundant if it is highly correlated with other features. Hence, the

Figure 1. Feature selection taxonomy



search for a good feature subset involves finding those features that are highly correlated with the decision feature(s), but are uncorrelated with each other.

A taxonomy of feature selection approaches can be seen in Figure 1. Given a feature set of size n, the task of any FS method can be seen as a search for an ''optimal'' feature subset through the competing 2^n candidate subsets. The definition of what an optimal subset is may vary depending on the problem to be solved. Although an exhaustive method may be used for this purpose in theory, this is quite impractical for most datasets. Usually FS algorithms involve heuristic or random search strategies in an attempt to avoid this prohibitive complexity.

Determining subset optimality is a challenging problem. There is always a trade-off in non-exhaustive techniques between subset minimality and subset suitability - the task is to decide which of these must suffer in order to benefit the other. For some domains (particularly where it is costly or impractical to monitor many features, such as complex systems monitoring (Shen & Jensen, 2004)), it is much more desirable to have a smaller, less accurate feature subset. In other areas it may be the case that the modeling accuracy (e.g. the classification rate) using the selected features must be extremely high, at the expense of a non-minimal set of features, such as web content categorization (Jensen & Shen, 2004b).

MAIN FOCUS

The work on rough set theory offers an alternative, and formal, methodology that can be employed to reduce the dimensionality of datasets, as a preprocessing step to assist any chosen method for learning from data. It helps select the most information rich features in a dataset, without transforming the data, 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, 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 rough set theory can not only 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.

Rough Set Theory

Rough set theory (RST) has been used as a tool to discover data dependencies and to reduce the number of attributes contained in a dataset using the data alone, requiring no additional information (Pawlak, 1991; Polkowski, 2002; Skowron et al., 2002). Over the past ten years, RST has become a topic of great interest to researchers and has been applied to many domains. Indeed, since its invention, this theory has been successfully utilized to devise mathematically sound and often, computationally efficient techniques for addressing problems such as hidden pattern discovery from data, data reduction, data significance evaluation, decision rule generation, and data-driven inference interpretation (Pawlak, 2003). Given a dataset with discretized attribute values, it is possible to find a subset (termed a *reduct*) of the original attributes using RST that are the most informative; all other attributes can be removed from the dataset with minimal information loss.

The rough set itself is the approximation of a vague concept (set) by a pair of precise concepts, called lower and upper approximations, which are a classification of the domain of interest into disjoint categories. The lower approximation is a description of the domain objects which are known with certainty to belong to the subset of interest, whereas the upper approximation is a description of the objects which possibly belong to the subset. The approximations are constructed with regard to a particular subset of features.

Rough set theory possesses many features in common (to a certain extent) with the Dempster-Shafer theory of evidence (Skowron & Grzymala-Busse, 1994) and fuzzy set theory (Wygralak, 1989). It works by making use of the granularity structure of the data only. This is a major difference when compared with Dempster-Shafer theory and fuzzy set theory, which require probability assignments and membership values, respectively. However, this does not mean that no model assumptions are made. In fact, by using only the given