Chapter 1

Introduction

1.1 Motivation

Technological advancement and explosive growth of electronically stored information in the area of computing has lead to the production of huge amount of structured as well as unstructured data. Millions of data is generated in multiple scenarios including weather, census, health care, government, social networking, production, business, and scientific research. They require new tools in order to be analyzed and processed so as to enable the extraction of useful information. Knowledge discovery from database (KDD) [5, 40, 45, 127] is a process of modelling databases or datasets to extract useful information from huge data repositories. The extracted information or knowledge is thereby novel, understandable, potentially useful and valid. Conventional methods for knowledge retrieval involves a lot of human intervention making the entire process too labour intensive, time consuming and highly expensive. Complexity of data further complicates the task of manual analysis making the process impractical for most applications. This necessitates the need for having an efficient system that can automatically extract informatton. KDD is a step forward.

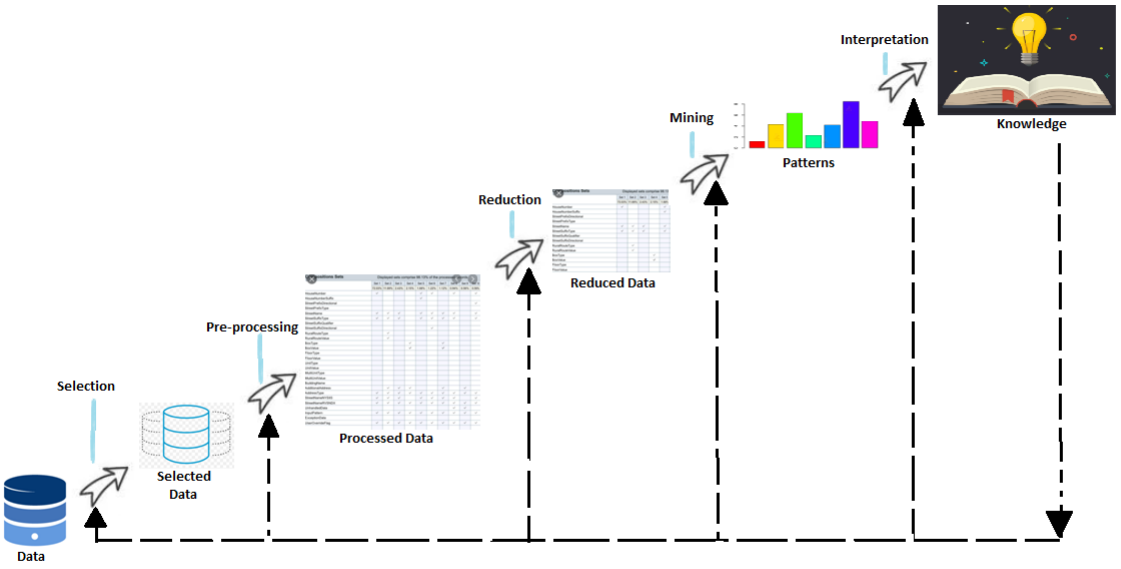

FIGURE 1.1: The knowledge discovery process

The knowledge discovery from databases can be identified as a process comprising the number of subtasks, as depicted in Figure 1.1:

1. **Dataset Selection**: The underlying dataset pertaining to the application is determined. The dataset can thus be selected or created by clubing various existing datasets.

2. **Data Preprocessing**: The selected dataset is undergone through a phase of data preprocessing to improve the quality of data. It comprises of tasks including data cleaning for removal of noise and inconsistencies arising in the data, missing value imputation, and outlier elimination. This stage of KDD refines datset quality so that the performance of subsequent stages of KDD is not diminished.

3. **Data Reduction**: This phase reduces the size of dataset to eliminate useless data that may mislead the entire task. The complexity of problem is also reduced if only relevant data is employed for further processing.

4. **Data Mining**: This phase of KDD is the most important stage that extracts the hidden patterns available in the dataset. Many factors govern the selection of algorithms, parameters and validation strategy.

5. **Evaluation/Interpretation**: The validity and usefulness of the identified knowledge must be reckoned based on some interestingness measures. This discovered knowledge can then be utilized for target application.

All the above phases of KDD are interconnected. The focus of this study is on third phase of KDD namely data reduction as loss can be significantly reduced if the supplied data contains only useful or relevant information. The dataset size can be reduced by removing irrelevant data employing suitable methods. The high dimensionality of the data can be reduced either by changing the underlying semantics of features or by selecting the subset of features based on subset's quality. Feature selection is an essential step as it enhances the undertanding of underlying information. Various factors like noise, missing values, outliers, etc affect the data reduction or feature selection task negatively. The subsequent sections will discuss some of the important terms related to each one of the above factors along with giving insight into feature selection stage.

1.2 Feature Selection

Features are also known as attributes. These features may have numerical, categorical, nominal or ordinal data type. The collection of all the features concerning a particular application forms a dataset. The dataset may be high dimensional comprising of huge number of features. Not all the features might be relevant to the problem at hand. Further, the high dimensional dataset may overfit the classifier

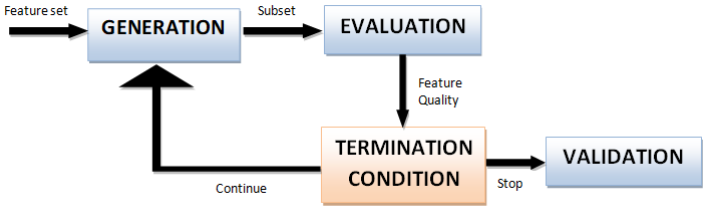

FIGURE 1.2: Feature Selection

degrading the performance because of noise and irrelevant features. The redundancy present among features may worsen the condition. To overcome such curse of dimensionality [8], feature selection comes as a remedial tool. Feature selection stage of KDD selects a subset of features based on certain evaluation criterion. The selected feature subset can thereby be the reflection of entire dataset. The task of feature selection can be thought of as comprising of two subtasks namely search strategy and feature quality evaluation (as shown in figure 1.2).

The number of feature subset that can be formed with N number of features is $2^N - 1$, a huge number of combinations to choose from. Therefore, an effective search strategy [14] is required that can guide in finding optimal subset without doing exhaustive search. Some of the heuristic search methods are:

- Forward selection: It begins with an empty set and adds features one by one until some termination condition is reached.
- Backward selection: Starting with entire feature set, it removes irrelevant features iteratively until convergence. A combination of forward and backward selection can be made such that it selects the most relevant feature and eliminates the worst feature concurrently.

- Random selection: Features, corresponding to this strategy are selected randomly according to certain criterion. Bio inspired search heuristic including particle swarm, ant colony, etc fall under this category.

The above defined search heuristic did not guarantee to generate optimal feature subset.

Once the feature subset is selected, it must be evaluated using some attribute evaluation measures like information gain, dependency, distance, consistency, classification accuracy, etc. Information gain measures the feature's quality based on uncertainty function. The two, uncertainty and information gain, being inversely proportional to each other. Distance indicate the discriminative ability of the feature, how well a feature is able to distinguish between different classes. The connection between two features is calculated by dependency measure. Consistency measure checks whether the consistency obtained by feature set is same as that of entire dataset. Based on these attribute evaluation measures, feature selection can be categorized as:

- Filter Approach: The quality of feature subset is evaluated based on certain criteria that depends on intrinsic properties of features itself (like information gain, dependency, etc) without requiring any feedback from classifier (as shown in fugure 1.3).

- Wrapper Approach: It evaluates the feature quality based on classification accuracy. Since, the classifier performance is employed for evaluating feature subset, a more accurate and high quality subset is obtained than filter approach. However, wrapper approach is computationally expensive becuase of feedback from the classifier (as shown in figure 1.4).

- Embedded Approach: It is the hybrid of both the above approaches thereby overcoming the drawback of both the approaches. It selects the feature subset

FIGURE 1.3: Filter approach to Feature Selection

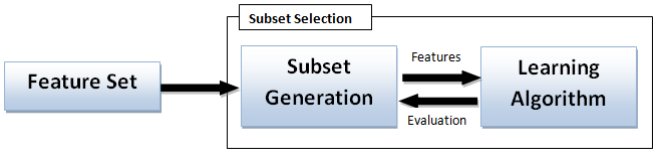

FIGURE 1.4: Wrapper approach to Feature Selection

based on feedback from the classifier and evaluates it using filter approach as well. It is therefore less time consuming than wrapper approach.

The entire process of selecting features can be done iteratively by forming a feature subset using search heuristic and then evaluating the subset quality until certain convergence criteria is reached. The choice of various algorithms and measures depends on the nature of dataset whether class labels are present or not i.e. supervised or unsupervised. Further, other factors like missing values and outliers must be handled.

- Supervised Dataset: The presence of the class label in the dataset i.e. supervised data may change the nature of algorithm employed for feature selection. A feature might be irrelevant and/or redundant depending to whether it is associated or adds something new to target class or assist in differentiating among classes.

- Unsupervised Dataset: The absence of class label i.e unsupervised domain will change the type of feature selection algorithm. Now, the relevant and non-redundant features must be chosen based on whether they provide the same information as that provided by original dataset or not.

- Class Imbalance: The variation in number of different classes may bias the classifier performance towards majority class. So, suitable balancing algorithm must be applied before feature selection.

- Missing Values: Lack of proper communication, instrument failure etc leads to presence of missing values in the dataset gathered. Hence, it is necessary to preprocess the datasets before analysing the performance of learning algorithms. Handling missing values [28, 55, 94, 113, 134, 162] is an important aspects that must be paid heed. Missing values can be handled in the following ways:

  - Ignoring the instance or sample: The instance containing missing value is ignored in further experiemntation. However, important information might be lost in such cases.
  - Filling the missing value: Replace the missing values with some common constant (like 0), statistical measures (like mean, median, mode). Such methods may bias the data.

  Therefore, the techniques for estimating/imputing missing values that can provide the actual representation of the data will be very helpful.

- Outlier Detection or Instance Selection: Apart from removing irrelevant and/or redundant features i.e handling curse of dimensionality, a huge volume of data instance may also be problematic and need to be analysed. The sheer volume of instances might be containing misleading or conflicting information. Clustering can be used to find groups and thereby eliminate outliers or spurious samples that might be degrading the model performance.

1.2.1 Benefits and Application of Feature Selection

Various potential benefits of applying feature selection are:

1. The data can be visualized conveniently if fewer features i.e. reduced representation are present and thereby complexity of learning models is reduced.

2. The storage space required will be reduced. This will be very beneficial for many applications like medical domain, etc.

3. The complexity and hence the time consumption will be decreased if dataset will be having low dimensionality.

4. Irrelevant and redundant features hamper the performance of learning algorithm. Removal of such noisy and misleading features enhances the prediction ability of the model.

The potential benefit of feature selection has further increased its applicability. It has got wide applications in fields namely image recognition [68, 132], text categorization, bioinformatics [58, 68, 75], etc (as shown in Figure 1.5). The most common issue observed is peaking where classification performance is decreased with increase in number of features after a peak. In melenoma diagnosis [58], only an accuracy between 65% to 85% is observed for distinguishing malignant from non-malignant melenomas which can be increased to as high as 95% by applying feature selection algorithm for skin tumour recognition. A huge number of features are generated in complex applications like industrial plants, many of which are redundant. An efficient feature selection technique will make the entire model more accurate and reliable [73, 78, 122, 129]. Dimensionality reduction is successfully applied in the field of text clustering to group together similar documents [29, 88]. In bioinformatics domain, features are selected effectively to reduce the size of dataset for

differentiating among healthy and cancer patients [163, 164].

The existing feature selection algorithms require user to supply value of parameters like size of feature subsets, noise levels, thresholds for stopping criteria, etc. This user intervention is a major drawback of feature selection algorithms. Rough set theory has proved to be successful for applying feature selection. It has been effectively applied for dimensionality reduction since last two decades. However, the applicability of rough set theory is only limited to crisp valued dataset. The real valued dataset need to discretised to attain crisp values for applying rough set theory. This leads to loss of information [11] as for example two real values like -1 and -400 would be mapped to same class "Negative" but -400 is much more negative than -1. Hence, this poses the need for an effective technique that can help in data reduction for real valued datasets without requiring human intervention. Fuzzy set theory is a solution that can handle real values by assigning degree of membership to each real value thereby handling the vagueness and uncertainty arising in the dataset. The extension of rough set theory i.e. fuzzy rough set theory provides greator degree of flexibility further by assigning a degree of membership in the range $[0, 1]$. It overcomes the limitations of rough set theory.

1.3 Thesis Structure

The rest of the thesis is organised as follows:

- **Chapter 2:** *Background*: Mathematical background of various theories like rough set theory, fuzzy set, etc used for feature selection is discussed in this chapter. Along with introducing these theories, the methodologies for employing it in feature selection is also described. The corresponding literature survey is also discussed.

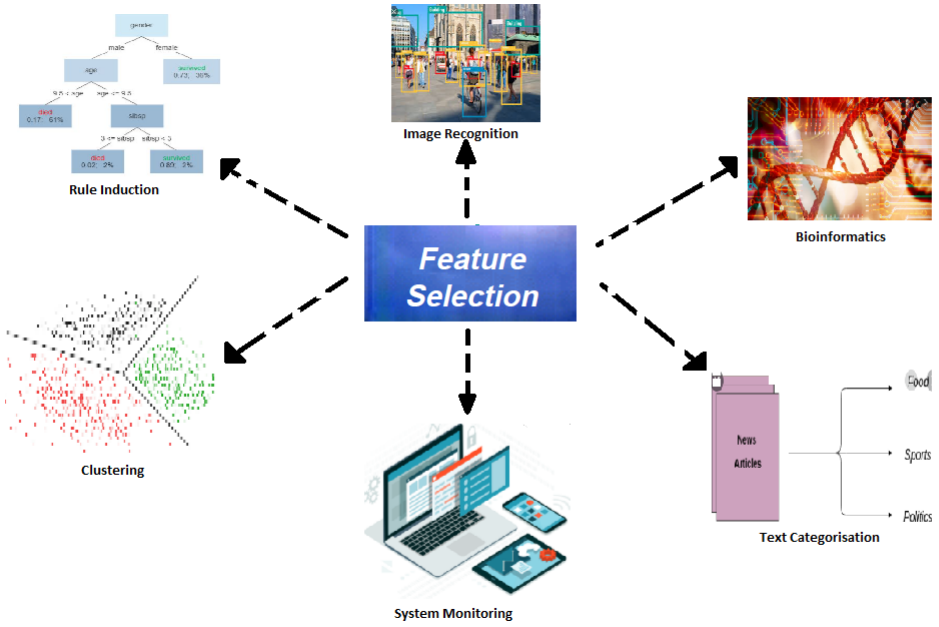

FIGURE 1.5: Various applications of Feature Selection

- **Chapter 3:** *Feature Selection Model and its Application*: In this chapter, various intuitionistic fuzzy rough set models are discussed with its applicaion in feature selection. Divergence based intuitionistic fuzzy rough set model is proposed and corresponding properties of lower and upper approximations are laid and proved. The resultant feature selection is applied for prediction of anti-tubercular peptides for tuberculosis treatment. Similarly, k-mean based intuitionistic fuzzy rough set based feature selection model is proposed to nullify the impact of noisy sample objects and is thereby utilized for prediction of aptamer-protein interacting pairs. The contents of this chapter is published in [65].

- **Chapter 4:** *A Fitting Model for Feature Selection*: A fitting intuitionistic fuzzy rough set model is introduced in this chapter in order to cope with uncertainty, vagueness arising in the datasets in much better way. This model fits data well and avoids misclassification properly. Firstly, Intuitionistic fuzzy

decision of an object is established using neighborhood concept. Then, intuitionistic fuzzy lower and upper approximations are introduced using intuitionistic fuzzy decision along with parameterized intuitionistic fuzzy granule and thereby dependency function is framed using these values. The content of this chapter is published in [66].

- **Chapter 5:** *Feature Selection Model for Incomplete Data*: Missing value imputation and feature selection is an efficient technique to overcome curse of dimensionality in incomplete datasets. Fuzzy rough set based approaches provide a handful of solutions for further dealing with vagueness and uncertainty available in the data. This chapter introduces the notion of imputing missing values followed by feature selection utilizing fuzzy rough set based approaches. The idea of missing value estimation and instance ignorance are combined for fuzzy rough missing value imputation employing only correlated features followed by feature selection with a search heuristic. The experimental evaluation on benchmark datasets demonstrates the applicability and robustness of the proposed work.

- **Chapter 6:** *Bireduct Model and its Application*: This chapter introduces the notion of bireducts in intuitionistic fuzzy framework that can be used for simultaneous reduction of instances and features. A robust lower approximation formula is employed along with laying the foundation for variants of instance selection technique. The experimental evaluation on benchmark datasets demonstrates the applicability and robustness of the proposed bireducts. It significantly reduces data size both in terms of instances and features whilst maintaining high performances. Further, the model is applied in the challenging domain of cancer treatment by enhancing the prediction performance of anti-angiogenic peptides [67].

- **Chapter 7:** *Unsupervised Feature Selection*: Tremendous amount of data is generated everyday from various sources that may or may not contain decision class. Manual annotation may not be feasible owing to large data size. High dimensionality of most of the datasets pose an additional challenge making the label assignment task cumbersome because of presence of redundant and/or irrelevant features. This prompts the need for feature selection in unsupervised domain. In the presence of decision class, the features are selected based on function so as to optimize parameters obtained either from or related to decision labels. However, the lack of decision class makes the feature selection task difficult. In this chapter, two unsupervised feature selection technique based on fuzzy rough sets using earthworm search strategy are proposed. The robustness has been demonstrated using experimental evaluation.

****