Chapter 1

Introduction

1.1 Motivation

We are living in a data driven era, in which digital data size usable in the world is continuously enlarging with the enhancement of computer as well as database technology. In the present scenario, due to advancement of internet-based technology, all business organizations regularly acquire data on millions of observations across diverse subjects, brands, at regular time periods, predictor variables and storage locations. Every day quintillions bytes of data are recorded at diverse nodal points like data pertaining to various banking and business transactions, bio-genetic informations in health services [28], enormous amount of statistical data regarding mass population and satellite data information of global and regional climate changes. New tools are always a requirement to analyse and process this large volume data so as to enable the extraction of useful information from the entire information system. This extracted information is the source of knowledge. Knowledge discovery in databases (KDD) [24],[87] is an exploratory and automatic analysis and modelling of large volume data repositories. KDD is the suitable and organized process of identifying novel, useful, understandable, and valid patterns from large and complex information systems [17, 58]. The abundance of data available today and their accessibility makes knowledge discovery a matter of considerable, necessity, and importance. The KDD process can be divided into the following stages:

1.1.1 Data generation

From an individual domain there is creation or acquisition of an objective dataset. For that purpose a merger of various existing datasets are involved to obtain an appropriate example set.

1.1.2 Data cleaning

This step includes many tasks to create processed data, such as discretization of attributes, noise removal, and missing value imputation, etc. The key issue is to enhance the overall property of any information that may be discovered from the information system.

1.1.3 Data reduction

Datasets dimensions are increasing in twofold, i.e. number of data points or instances and number of features. Most datasets usually consist of certain amount of redundancy that does not aid in knowledge discovery and may actually mislead the entire process. The main objective of this phase is to find valuable features to represent the entire data and eliminate non-relevant as well as redundant features and similar data points. This step also aids in preserving the time, storage and cost during the data mining process and also improves the interpretability of data.

1.1.4 Data mining

Data mining is defined as the process of discovering interesting patterns and extraction of knowledge from large volumes of data. Data mining is a collection of practices used in an automated method to exhaustively explore and carry to the surface complex relationships in large amount datasets. The data sources can be databases, the Web, data warehouses, other information repositories, or data that are fed into the system dynamically.

1.1.5 Interpretation/evaluation

After the discovery of knowledge, it is evaluated with respect to novelty, validity, simplicity, and usefulness. This may involve repeating some of the former steps.The entire process is summarized in Figure 1.1, where the dimensionality reduction (DR) phase is a preprocessing stage in the complete system.

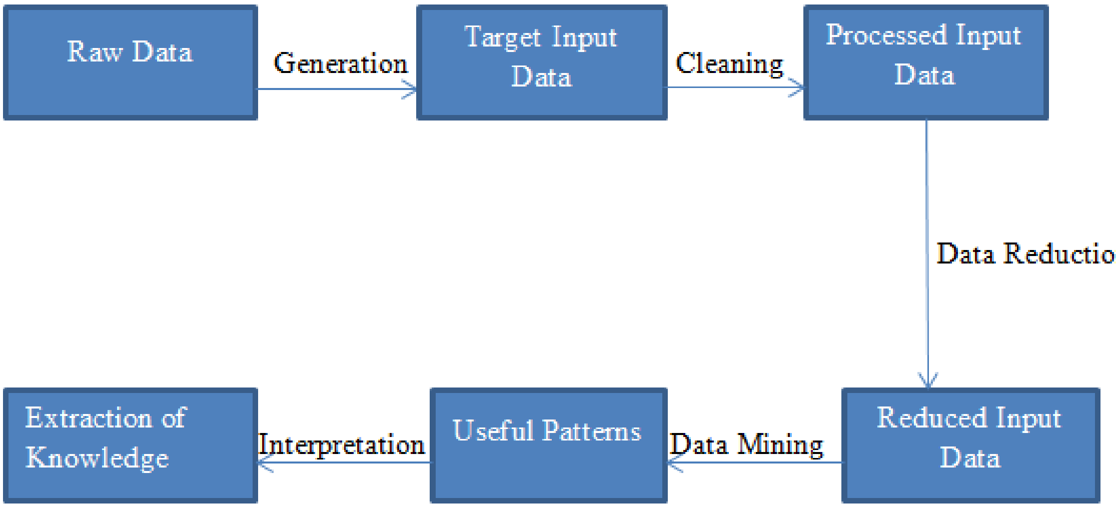


Figure 1.1: Steps of Knowledge Discovery

Curse of Dimensionality:

Data generation is stirring at a record rate. This increase can be seen in almost fields of human activity right from the data created on a daily basis, such as bank transactions, business tractions, telephone calls, etc. to more technical and complex data, including molecular datasets, medical records, astronomical data, genome data, etc. These datasets may comprise lot of information convenient but still undiscovered. This increase in data can be one of the twofold, i.e. either in number of samples/instances or number of features that are recorded and calculated. As a result of this increase, many real world applications requisite to process datasets with hundreds and thousands of features or attributes. Some of such datasets are also publicly available at UCI. The significant rise in the number of dimensions in datasets leads to the reason called curse of dimensionality. It results due to the exponential increase in volume related with the addition of extra

dimensions to a space. Dimension reduction is applied as a preprocessing step in KDD. The original feature space is mapped onto a new space, reduced dimensionality space, and the samples are characterized in that new space . Normally datasets comprise a number of irrelevant or non-predictive and redundant information which requires to be removed before any further processing can be performed on these datasets. For example, it is more effective when dimension reduction has to be performed first in order to derive complex classification rules. This step enhances the accuracy of resulting classification rather not only improvement in performance and also adds more comprehensibility in available rules. There are various techniques to perform dimensionality reduction, some of such techniques destroy the underlying semantics of data, i.e. original meaning of data, which may make them undesirable to perform many real-world applications, wheareas some of these techniques preserves the semantics of data and enhances the interpretability of data. Dimensionality reduction can be categorized as follows:
Depending on the necessities of future knowledge discovery in database, by using appropriate methods the high dimensionality of large volume database can be reduced. These techniques can be categorized into two parts: those that transform the original meaning of the data features and those that preserve the semantics of the features. Feature selection techniques belong to the second category, where a subset of the underlying features present in a given dataset is selected based on a subset evaluation function. In the process of knowledge discovery, feature selection techniques are particularly needed as these facilitate the interpretability of the resultant knowledge.

1.1.6 Transformation Based Techniques

Common throughout the DR literature are approaches that reduce dimensionality but in the process irreversibly transform the descriptive dataset features. These methods are employed in situations where the semantics of the original dataset are not needed by any future process. This section briefly discusses several such popular techniques, which are separated into two categories: those methods that are linear and those that are nonlinear.

1.1.6.1 Linear Methods

Linear methods of dimensionality reduction have been well developed over the years and include techniques such as Principal Component Analysis (PCA), Multidimensional Scaling, and Projection Pursuit.These techniques are used to determine the Euclidean structure of internal relationships of a dataset. However, when such relationships are of a higher dimensionality, these methods generally fail to detect this. This problem is not too restrictive as for many applications linear dimensionality reduction is all that is needed.

1.1.6.2 Non-linear Methods

Previous methods are useful for reducing dimensionality of linear datasets, their utility fails for nonlinear data. Given a dataset containing nonlinear relationships, these methods detect only the Euclidean structure. This brought about the need for methods that can effectively handle nonlinearity. The first attempts were extensions to the original PCA process, either by clustering data initially and performing PCA within clusters, or by greedy optimization processes. Both suffer from problems brought on as a result of simply attempting to extend linear PCA. This motivates the development of techniques, such as Isomap and LLE, designed to suitably and successfully handle nonlinearity.

In the following section, some of the important definitions related to feature selection are discussed.

1.2 Preliminaries

1.2.1 Feature

A feature [18] represents a characteristic or an attribute of an object or data point, where an object denotes an entity possessing a physical existence. In a dataset, data can be presented in the form of a matrix of “i” rows and “j” columns. Each row denotes a record or object or instance or data point, whereas each column denotes a feature or attribute. So, in a dataset, each data point is expressed in the form of a collection of attributes or

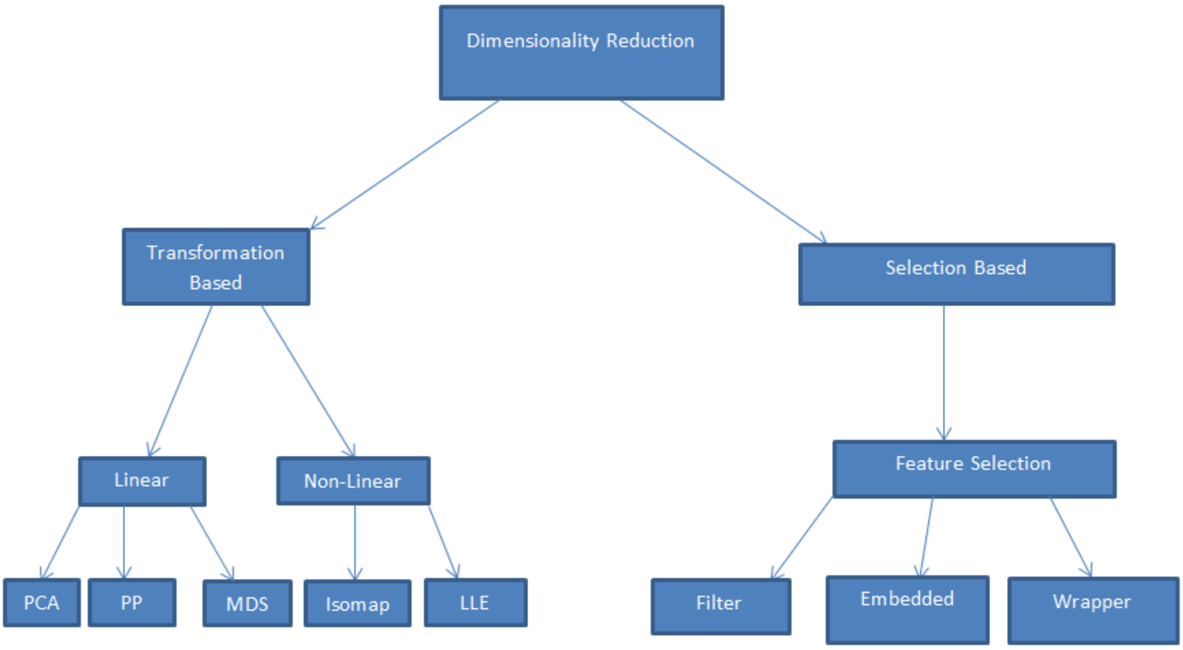
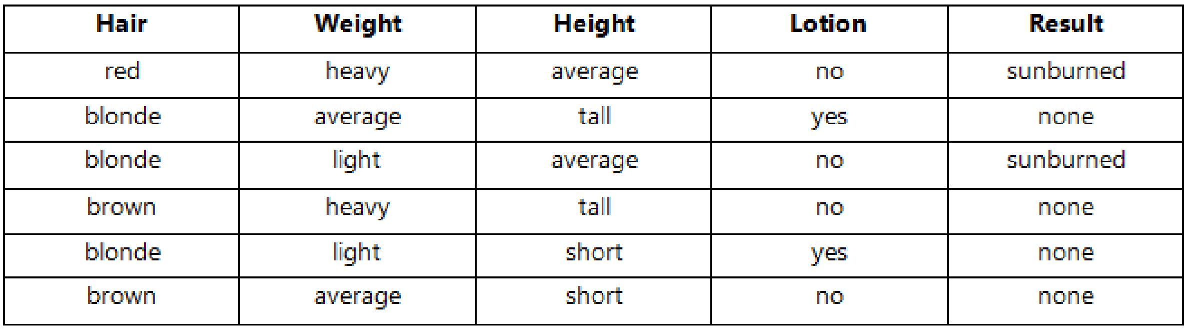


Figure 1.2: Categorization of dimensionality reduction techniques

features. An example dataset is given in Table 1.1, where five features are given, one of which is of class "Result". The class feature has two values: sunburned or none. Each row in the table is an instance. All features are contained of discrete values.

Table 1.1: Example dataset



Hair	Weight	Height	Lotion	Result
red	heavy	average	no	sunburned
blonde	average	tall	yes	none
blonde	light	average	no	sunburned
brown	heavy	tall	no	none
blonde	light	short	yes	none
brown	average	short	no	none

On the basis of nature or domain of data and type of the values a feature or attribute may contain, features can be partitioned into two types:

1.2.2 Numerical

Numerical features are defined as those attributes features that contain numbers as their values. For example, height of a person can be given as 8 ft. and 4 in., is a numerical value.

On the basis of range of values, numerical features can be either discrete or continuous. Numerical features can be further classified into two types:

1.2.2.1 Interval-scaled

Features, where the difference between two values is always meaningful, should be illustrated as Interval-scale features. For example, the difference between 80 and 60 is the same as the difference between 60 and 40.

1.2.2.2 Ratio scaled

These features have all the properties of interval-scaled features with additional characteristics of defined ratios for data analysis. For example, for attribute age, it is obvious that someone who is 45 years old is thrice as old as someone who is 15 years old.

1.2.3 Categorical Attributes

Categorical attributes use words to represent domain values. For example, gender can be presented by two symbols “M” and “F” or “Male” and “Female”. These attributes can further partitioned into two parts as follows:

1.2.3.1 Nominal

Nominal features are the ones where order does not matter. For example, the feature “gender” is said to be a nominal feature as the domain values “M” and “F” do not involve any order.

1.2.3.2 Ordinal

Ordinal features of data points are the features where order matters, i.e. both inequality and equality can be involved. For example, “qualification” feature can be represented as an example of ordinal type as one can involve “equal to”, “less than” and “greater than” operators for comparison purpose. Figure 1.2 represents categorization of feature types.

1.3 Feature Selection

Feature selection [22; 27; 30; 36; 37; 42; 45; 47; 48; 51; 62; 71; 75; 79; 84; 91] is a preprocessing steps in KDD that chooses an optimal subset of available features according to a certain criterion. The criterion considers the details of measuring feature subsets. The objective of feature selection or attribute reduction influences the selection of a particular criterion. An optimal subset can be a minimal subset, for which the values of other evaluation metrics being equal, it can be a subset that gives the best estimation of overall predictive accuracy. In some cases, if the number of features are given (as in data visualization and, projection, this number is usually 2 or 3), one needs to obtain a subset with the specified number that satisfies the criterion best (such as degree of dependency, information gain, consistency, etc.). An optimal feature subset selected or extracted by a dimensionality reduction method is always relative to a certain feature evaluation criterion. In general, different criteria may lead to different optimal feature subsets. However, every criterion tries to measure the discriminating ability of a feature or a subset of features to distinguish different class labels. Uncertainty is one of the main problems in real life data analysis. Some of the sources of this uncertainty include incompleteness and vagueness in class definitions. In this background, the possibility concept introduced by rough set theory has gained popularity in modelling and propagating uncertainty. It has been applied to reasoning with uncertainty, fuzzy rule extraction and modelling, classification, clustering, and feature selection. Feature selection [14] methods are utilized for identification and removal of the unwanted, redundant and irrelevant attributes from dataset which has no contribution towards the accuracy of any predictive model. For reduced complexity of the model, least number of attributes are desired. However, a brute force approach cannot be applied to feature selection, since, the number of feature subsets with m features from a collection of N total features is ${}^{n}C_{m}$.

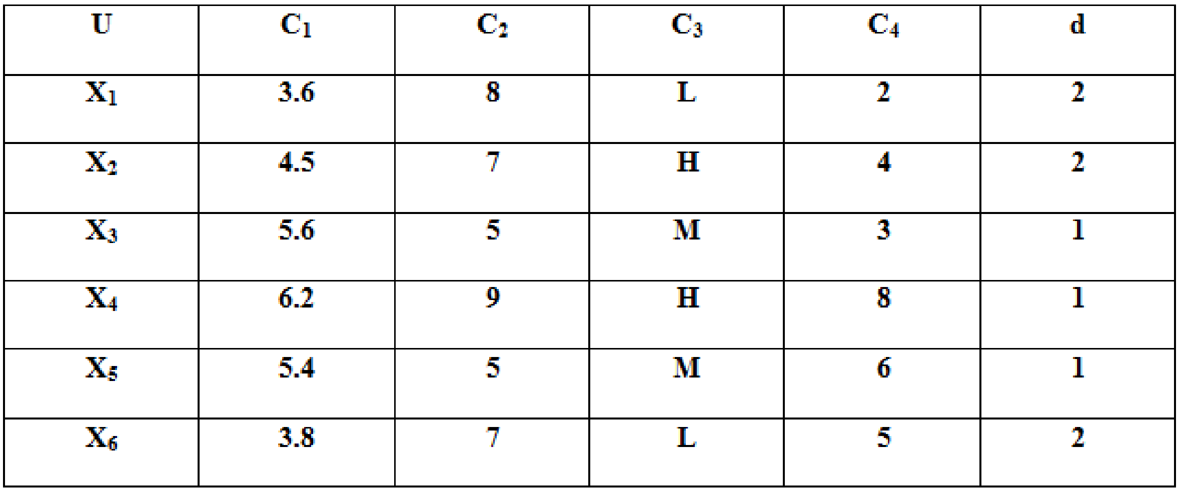
Based on the nature of existing data, feature selection process can be classified into three categories as follows:

1.3.1 Supervised Feature Selection

The majority of real-world classification problems involve supervised learning where the original class probabilities and class-conditional probabilities are not known, and every instance or tuple is related to a class label. However, we frequently face the data having hundred and thousands of features. The existence of noisy, irrelevant, and redundant features is the key issue that one has to usually face in the classification process. A predictive feature is neither redundant nor irrelevant to the target notion whereas a non-predictive feature is not directly related to the target concept, but affects the learning of the classifiers, and a redundant feature does not improve anything new to the target idea [14]. Normally it is hard to filter out the significant features, particularly when the data is enormous. Therefore, the machine learning algorithms perfomance and accuracy is affected. So, it is required to preprocess data before feeding to classifiers. Supervised feature selection plays its vital role to select minimum and predictive features. In supervised feature selection, the class labels are already known along with attributes or features where each tuple or instance belongs to an already identified class label, and feature selection algorithm chooses features of the underlying features based on some specific criteria. The selected features are then given as input to the machine learning algorithms.

Table 1.2 represents an example of supervised dataset. In Table 1.2, C_1, C_2, C_3, C_4 are conditional features and “d” is decision class or class label.

Table 1.2: A sample of supervised dataset

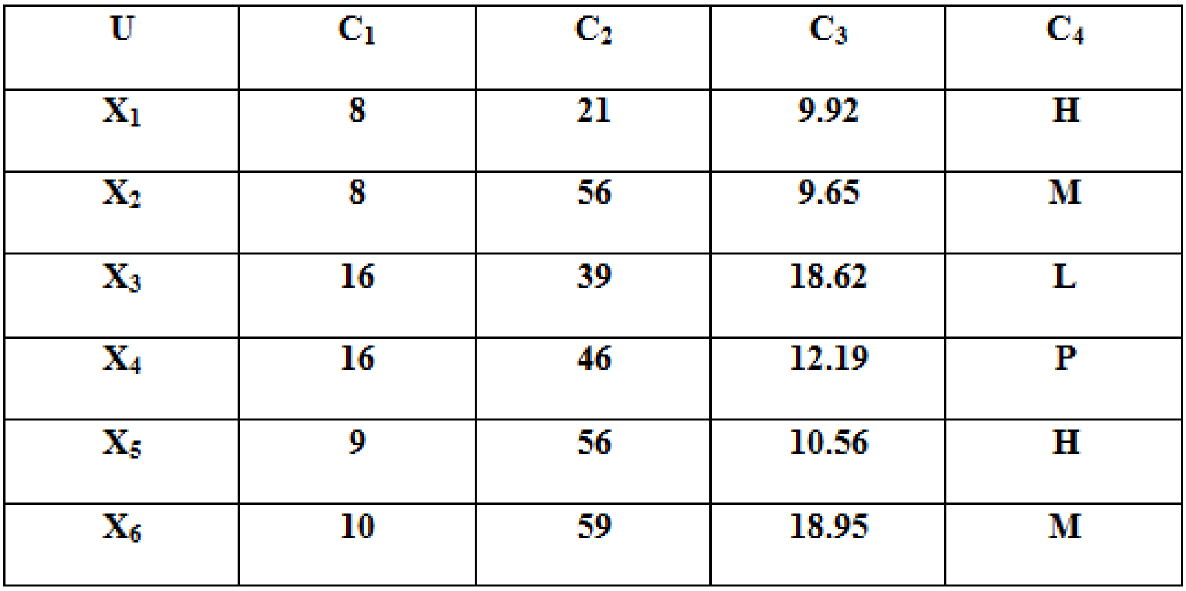


U	C_1	C_2	C_3	C_4	d
X_1	3.6	8	L	2	2
X_2	4.5	7	H	4	2
X_3	5.6	5	M	3	1
X_4	6.2	9	H	8	1
X_5	5.4	5	M	6	1
X_6	3.8	7	L	5	2

1.3.2 Unsupervised Feature Selection

It is not essential that classification information is known all the time. In unsupervised learning, the features are given without any class label. The machine learning algorithms have to use only the existing information. So, a simple policy may be to create the clusters (a cluster is an assemblage of similar objects, analogous to a classification structure except that class labels are not available and hence only underlying features are applied to form clusters, whereas in generating classification structures, class labels are provided and used). Now, yet again the problem of irrelevant, noisy, and redundant data prohibits the usage of all of the attributes or features to feed to machine learning algorithms. Furthermore, eliminating such features is again a clumsy task for which manual filtration may not be promising. Again, we have to take benefit from the feature selection process. So, mutual criteria are to select those attributes or features that provide the same clustering structure as provided by the complete set of features. Table 1.3 shows an example of unsupervised dataset, where objects are characterized by the set of features C_1, C_2, C_3, C_4.

Table 1.3: A sample of unsupervised dataset

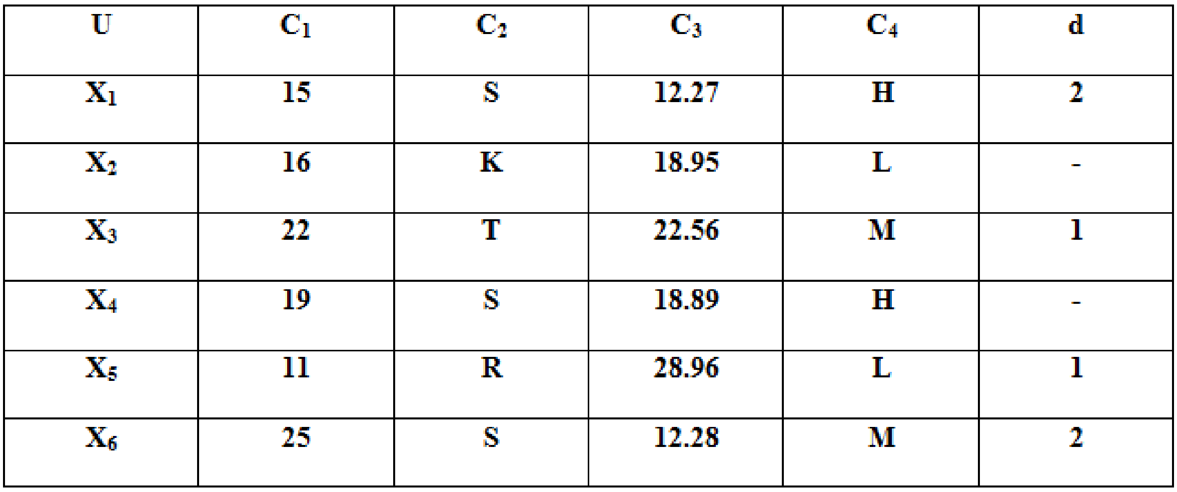


U	C_1	C_2	C_3	C_4
X_1	8	21	9.92	H
X_2	8	56	9.65	M
X_3	16	39	18.62	L
X_4	16	46	12.19	P
X_5	9	56	10.56	H
X_6	10	59	18.95	M

1.3.3 Semi-supervised feature selection

When a small number of instances are labelled, but the majority is not labeled, semi-supervised feature selection is designed to take advantage of both the large number of unlabeled instances and the labeled information. In other words, semi-supervised feature selection attempts to align locality-based separation and class-based separations. Since there are a large number of unlabeled data and a small number of labeled instances, it is reasonable to use unlabeled data to form some potential clusters and then employ labeled data to find those clusters that can achieve both locality-based and class-based separations. Table 1.4 represents an example of semi-supervised dataset, where features are C_1, C_2, C_3, C_4 and class label is d with missing values.

Table 1.4: A sample of semi-supervised dataset



U	C_1	C_2	C_3	C_4	d
X_1	15	S	12.27	H	2
X_2	16	K	18.95	L	-
X_3	22	T	22.56	M	1
X_4	19	S	18.89	H	-
X_5	11	R	28.96	L	1
X_6	25	S	12.28	M	2

Feature selection approaches can be classified into three categories based upon relationship between learning algorithm and selected features or on the basis of evaluation of selected feature subset.

1.3.4 Filter Methods

Filter-based feature selection methods use a statistical measure to assign a score to each feature. Based on the scores, the features are ranked and are either selected to be kept or removed from the dataset. These methods are often univariate and consider the features independently, or about the dependent variable. Filter methods [82] are the most straightforward technique for feature selection. In filter methods, the process of feature selection remains independent of any machine learning algorithm, i.e. no feedback from the learning algorithm is required. In effect, irrelevant features are filtered out prior to induction. Features or attributes are evaluated based on some specific criteria (Information gain, Chi-squared test, and correlation coefficient scores) by using the essential characteristics of the features. So, the classifier or machine learning algorithm has no control over the quality of the selected features. Filter methods can be further divided into two subcategories:

1.3.4.1 Attribute evaluation methods

In attribute evaluation methods, each individual feature is evaluated according to the selection criteria and each feature is ranked, after which a specific number of attributes or features are selected as the output.

1.3.4.2 Subset evaluation methods

The evaluation methods of feature subset on the other side evaluate the entire subset on the basis of definite criteria. Figure 1.3 shows the generic diagram of filter-based approach [45].

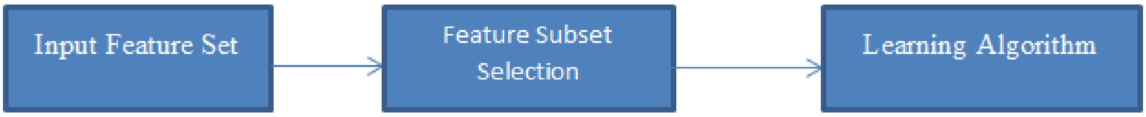


Figure 1.3: Generic model of filter-based feature selection

1.3.5 Wrapper Methods

Wrapper methods use a predictive model to assign scores to subsets of features, which are hence compared, to find the most informative feature subset. The search process may be methodical (e.g. best-first search), stochastic (e.g. random hill-climbing algorithm), or may use heuristics, like forward and backward passes to add and remove features (e.g. recursive feature elimination algorithm). Filter methods choose optimal attributes or features independent of any learning algorithm; however, quality and optimal feature subsets are usually dependent on the heuristics and generally biases of the machine learning algorithm, so, it should be related with and selected by using an underlying classification algorithm. This is the original concept of wrapper techniques [50] . So, in distinction with the filter methods, wrapper techniques do not select underlying features independent of the classification algorithm. Therefore, the feedback from the learning algorithm is applied to measure the quality of selected features and thus results in better quality and high performance of classifiers. It can be noticed that there is a two-way link among “feature subset search”, “evaluation” and “induction” processes, i.e. attributes or features are evaluated and consequently searched again based on the feedback from induction algorithm.

- Search feature subset from the available features set.
- Evaluate features based on induction algorithm (learning algorithm).
- Continue process until it obtains optimized feature subset.
  It is obvious that induction algorithm or classification algorithm is used as a black box where the selected feature subset from the available feature set is directed and the acknowledgement is accepted in the form of some quality measure, such as error rate (for example standard deviation).

1.3.6 Embedded Methods

Embedded methods find the features which contribute highly to the accuracy of the model. The most common type is regularization methods. Regularization methods are also known as penalization methods which introduce additional constraints into the optimization of a predictive algorithm which in fact bias the model toward lower complexity (fewer coefficients). Examples are Elastic Net, LASSO, and Ridge Regression. Embedded methods have a tendency to overcome the shortcomings of both filter and wrapper techniques. In Filter methods, features are evaluated independently from classification algorithm, while in wrapper approaches, features are evaluated by using feedback from classifier, which is computationally expensive as classification algorithm runs many times to select the reduct or optimal feature subset. In Embedded models,feature subsets are constructed as part of the classification algorithm, so that they can observe advantages of both wrapper method (feature subset evaluation is not independent of learning algorithm) and filter method (moreover selected features are evaluated on the basis of independent measures). The entire functionality of embedded method can be given as follows:

- Initialize attribute or feature subset (either empty or comprising all the features).

- Evaluate the subset based on independent measure.

  If it satisfies criteria more than current subset, this converts into the current subset.

- Evaluate the subset with respect to evaluation criteria specified by classification algorithm.

  If it again fulfils criteria more than current subset, this converts into the current subset.

- Repeat Step 2 to Step 3 until termination criteria are met. There are three categories of embedded techniques. The first are pruning techniques, in which initially the model is trained using the entire set of features and then features are eliminated gradually. Then the second models that provide built-in mechanism to accomplish feature selection process, and lastly there are regularization models that minimize fitting errors and concurrently eliminate the features by imposing the coefficients to be small or zero.

1.4 Feature Selection Criteria

The central concept of feature selection is the selection of potential feature subset. The potential feature subset is selected from the entire dataset based on some criteria [30]. These criteria are as follows:

1.4.1 Information Gain

When one receives the messages, then uncertainty is measured in terms of information. If the receiver is familiar of what is coming, his contemplation surprise level (uncertainty) is small; if he is not familiar of all what is coming, a reasonable postulation is that all messages have nearly equal probabilities to come, his anticipated surprise level is high. In the case of classification process, messages are known as classes. An information measure U is said to be as the uncertainty function regarding the true class, and is expressed as the larger values for U characterize higher levels of uncertainty. Information gain

can be defined in terms of "uncertainty". The maximum the uncertainty is available in the information system; the information gain will be the minimum in value. Let prior class probabilities be P(C_m), where $m = 1, 2, , d$, and if IG(K) is the information gain from feature K, then feature K will be better over another feature L only if IG(K) $>$ IG (L). If $E[X]$ represents expected value of X and "U" is an uncertainty function, P(C_m) is class C_m probability prior to considering feature "K" and P(C_m|K) denotes posterior probability of class C_m by considering the feature "K", the information gain will be:
$IG(K) = \Sigma_m U(P(C_m)) - E[\Sigma_m U(P(C_m|K))]$

So, the information gain can be demonstrated by difference of earlier uncertainty and uncertainty after considering feature K.

1.4.2 Distance

Distance measure describes the discrimination ability of feature, i.e. with how much powerfully a feature can discriminate objects among classes. A feature with higher discrimination ability is said to be better than the one with little discrimination ability. If C_i and C_j are two classes and K is any feature or attribute, then the distance measure D(K) will be defined by the difference of P(K|C_i) and P(K|C_j), i.e. the difference of probability of "K" when class is C_i and when class is C_j. K will be preferred over L if D(K) $>$ D(L). If the distance is the maximum, then the feature is preferred more. If P(K| C_i)=P(K|C_j), feature K cannot differentiate classes C_i and C_j.

1.4.3 Dependency

Rather than the convergence power or information gain, dependency measure describes how strongly two features are associated with each other. In other words dependency describes about how uniquely the values of a feature define values of other features. In supervised learning, it could be explained as the dependency of class label "C" on feature "X" while in unsupervised learning, it may be defined as the dependency of other features on the one under consideration. The features may be selected on which other features have high dependency value. If D(X) denotes the dependency of class C over feature X, then attribute or feature X will be preferred upon feature Y if D(X) > D(Y).

1.4.4 Consistency

One of the well-known criteria of feature selection is to choose those features that supply same class structure as supplied by entire feature set. The entire mechanism is called consistency, i.e. to choose the features along with the condition P(C| Full Set)= P(C|Subset). So, those features are selected that sustain same consistency as sustained by the entire dataset.

1.4.5 Classification Accuracy

The accuracy of classification algorithm is generally dependent on the classifier applied and is appropriate for wrapper-based methods. The goal is to select the features that give the best classification accuracy on the basis of feedback from the classifier. Though this measure produces quality features (as learning algorithm is also involved), but it has some drawbacks also, e.g. how to calculate accuracy by avoiding over fitting as noise in the data may lead to inaccurate accuracy measure. It is expensive to compute accuracy as classification algorithm takes time to learn from the information system.

In the feature selection process, one of the most important concepts is the way of selection of the features from the entire dataset. This process is called feature generation scheme.

1.5 Feature Generation Scheme

For any feature selection algorithm, the next feature subset generation is the key point, i.e. to select the members of feature subset for the next attempt if in the current attempt the selected feature subset does not provide appropriate solution. For this purpose, we have four basic feature subset generation schemes.

1.5.1 Forward Feature Generation

In forward feature generation scheme, one starts with an empty feature subset and features are added one by one until feature subset satisfies the required criteria. The maximum size of feature subset selected may be equal to the total number of attributes or features in the entire dataset.

1.5.2 Backward Feature Generation

Backward feature generation is opposite to forward generation scheme: in forward generation, one starts with an empty set and features are added one by one. In backward feature generation, on the other hand, one starts with full feature set and keep on eliminating the features one by one, until no further feature or attribute can be removed without influencing the specified criteria.

1.5.3 Random Feature Generation

There is another strategy known as random feature generation. In random feature generation apart from all of the above-mentioned strategies, one randomly selects features to be part of the entire solution. Features can be selected or skipped on any specified criteria. A simple process may be to select the features on the basis of random number generation mechanism.

1.6 Advantages of Feature Selection

Several advantages of feature selection may include:

- Data visualization: patterns and trends within the data can be visualized and recognized more easily by reducing the dimensionality of data. It is found to be important in scenarios where only a few features have a significant effect on data outcomes. Learning algorithms may not be able to discriminate these factors from the rest of the feature or attribute set, hence it produces quite complicated models, the interpretation of which is very expensive.

- The measurement and storage requirements: In various domains where cost and time-expense of taking the measurements for data attainment is significant, only fewer features or attributes are desirable due to the expense involved. Also, in domains where large datasets are encountered, and space becomes an issue, manipulated reduction in data size is required to enable storage requirements.

- Training and utilization times: Feature selection for big data results in smaller data size which in turn significantly enhances the running times of machine learning algorithms for both the training and classification steps.

- Prediction performance: Feature selection in several cases enhance overall accuracy of classifiers, through the removal of misleading, redundant and noisy features. When machine learning algorithms are trained on a full set of features may not be able to discern these features or attributes resulting in lesser accurate data for unseen samples.

1.7 Various applications of feature selection

Feature selection or attribute reduction has various applications in the fields of data mining, signal processing, image recognition, Bioinformatics, text categorization, systems monitoring, clustering data and rule induction [27; 37; 42; 91]. Peaking is a process commonly observed when classifiers are trained with a limited set of example sets. If the number of attributes or features in the data is increased, the classification rate of the classifier ceases to increase or in some cases decrease after a peak. For example, in melanoma diagnosis, for instance, the clinical accuracy of dermatologists in identifying malignant melanomas is only between 65% and 85%. With the application of feature selection algorithms, the accuracy of automated skin tumour recognition systems can go as high as 95%. One other application of Feature selection or attribute reduction is in the field of gene expression micro-arrays, a technology which can help analyze the expression levels of tens of thousands of genes in a single experiment. It is used to distinguish between cancerous and healthy cells. In such scenarios, feature selectors are used in reducing the size of the gene datasets, which otherwise are unsuitable for further processing. Structural as well as functional data from analysis of the human genome are increasing many fold in past years. Feature selectors are applied to reduce the size of these datasets, which would otherwise have been unsuitable for further processing. Other applications within bioinformatics include QSAR (Quantitative structure activity relationship), where the main objective is to form hypotheses relating chemical features of molecules to their molecular activity, and splice site prediction, where junctions between coding and non-coding regions of DNA are detected.

The most common technique to establish expressive and human readable representations of knowledge is the use of if-then production rules. Real-life problem domains are yet lacking generic and systemic expert rules for mapping feature patterns into their underlying classes. The rule induction process can be speed up using a suitable feature selection process by reducing rule complexity. Moreover, many inferential measurement systems are developed with the help of data based methodologies. The models are used to infer the value of target class label. This implies that inferential systems are heavily affected by the quality of the data used to develop their internal models. So, complex application problems, such as diagnosis of industrial plants and reliable monitoring , are likely to present huge number of features , many of which will be irrelevant and redundant for further processing. Additionally, it is found that there is a related cost with the measurement of these features. So, It is always useful to have an intelligent system capable of selecting the most relevant features needed to form an accurate and the most reliable model for the further processing. Feature selection applications in various fields can be given by the Figure 1.4.

In the crisp approaches of feature selection, the use of user-supplied information is the most essential part. This is itself a significant drawback as some feature selectors involve noise levels to be specified by the user beforehand. Some of the algorithms simply rank features leaving the user to select their own subset. Some times the users have to state how many features are to be selected and they have to provide a threshold that decides when the algorithm should terminate. However all of these require the user to make a

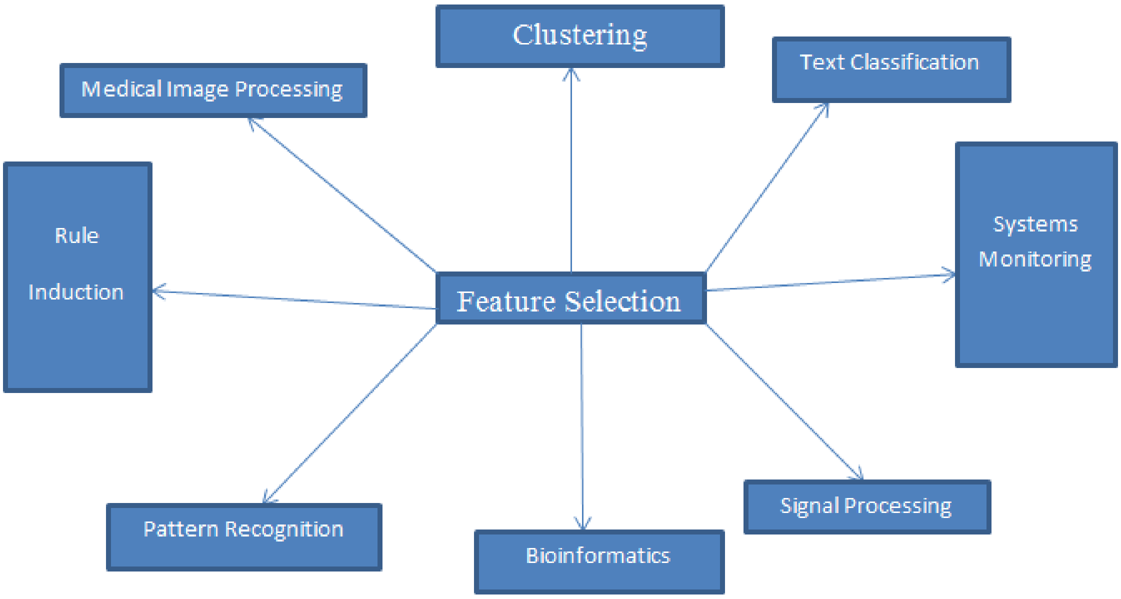


Figure 1.4: Various Applications areas of Feature Selection

judgement based on their own decision. The use of rough set theory (RST) [67; 68] to attain data reduction is one approach that has been proved successful. Over the last twenty years, rough set theory has become a topic of great interest to researchers and has been implemented to many domains (e.g. classification , systems monitoring , clustering , expert systems [97; 108]). This success is due 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.

  Given a dataset with discretized attribute values, it is possible to find a subset of the original attributes using RST that are the most informative (termed as reduct); all other attributes can be removed from the dataset with minimal information loss. This method tends to be a pre-processing step to reduce dataset dimensionality before some other action is performed (for example, induction of rules).

1.8 Rough Set Theory

Rough set theory [67; 68],[69] can be used to extract knowledge from a domain in a concise way which, even while reducing the amount of knowledge involved, can retain the information content. Discernibility is an important concept of RST which can be used for feature selection.

Definition 1.8.1 *A quadruple (U, A, V, f) is said to be an information system, where U (the universe of discourse) is a non-empty set of finite objects, A is a non-empty finite set of attributes, V is the set of attribute values and f is an information function from $U \to V$. If $A = C \cup D$ such that $C \cap D = \phi$, where C is the set of conditional attributes and D is the set of decision attributes, then (U, A, V, f) is known as decision system.*

For any $P \subseteq A$ there is an associated equivalence relation R_P:

$$R_P = \{(x, y) \in U^2 \mid \forall a \in P, a(x) = a(y)\} \tag{1.1}$$

If $(x, y) \in R_P$, then x and y are said to be indiscernible by attributes from P. $[x]_P$ denotes the equivalence classes of the P-indiscernibility relation. Let $X \subseteq U$; X can be approximated using the P-lower and P-upper approximations of X where the lower and upper approximations are defined as below:

$$R_P \downarrow X = \{x \in U \mid [x]_P \subseteq X\} \tag{1.2}$$

$$R_P \uparrow X = \{x \in U \mid [x]_P \cap X \neq \phi\} \tag{1.3}$$

The tuple $\langle R_P \downarrow X, R_P \uparrow X\rangle$ is called a rough set. In the given Figure 1.5, $\underline{R}X$ represents $R_P \downarrow X$ and $\overline{R}X$ represents $R_P \uparrow X$.

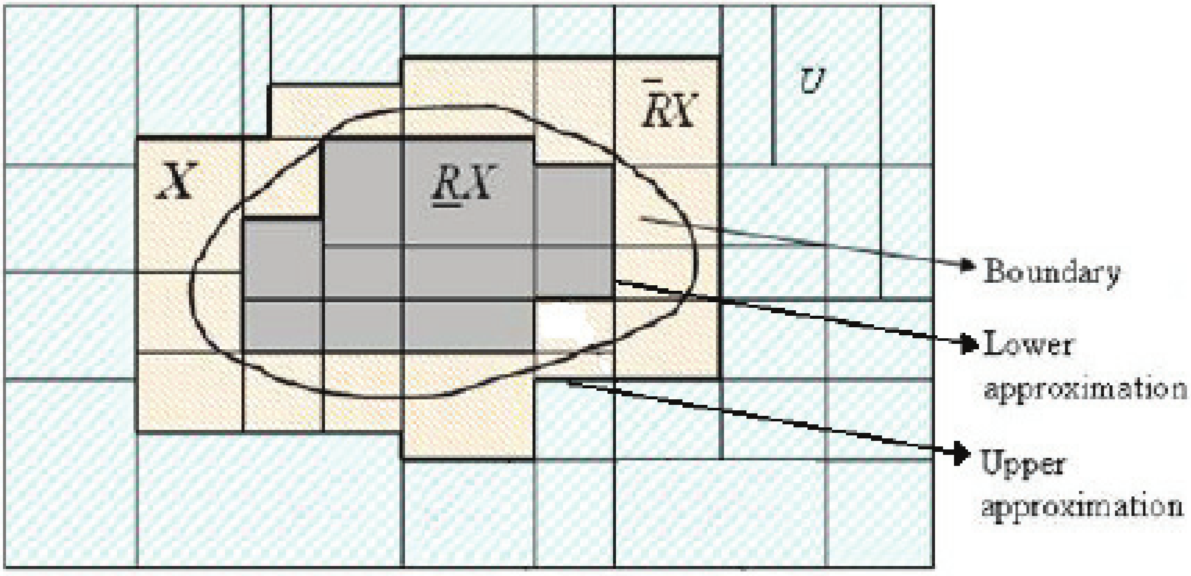


Figure 1.5: Graphical Representation of Rough Set

1.9 Rough Set based Feature Selection

Let [80] $(U, C \cup D, V, f)$ be a decision system with $Q \in D$ as an decision attribute. Its equivalence classes $[x]_{R_Q}$ are called decision classes. Given $P \subseteq C$, then P-positive region (POS_P) comprises those objects from U for which the values of P allow to predict the decision class clearly:

$$POS_P(Q) = \bigcup_{x \in X} R_P \downarrow [x]_{R_Q} \tag{1.4}$$

Indeed, if $x \in POS_P$, it means that whenever an object has the same values as x for the attributes in P, it will also belong to the same decision class as x. The predictive ability with respect to Q of the attributes in P is then measured by the following value (degree of dependency of Q on P):

$$\gamma_P(Q) = \frac{|POS_P(Q)|}{|U|} \tag{1.5}$$

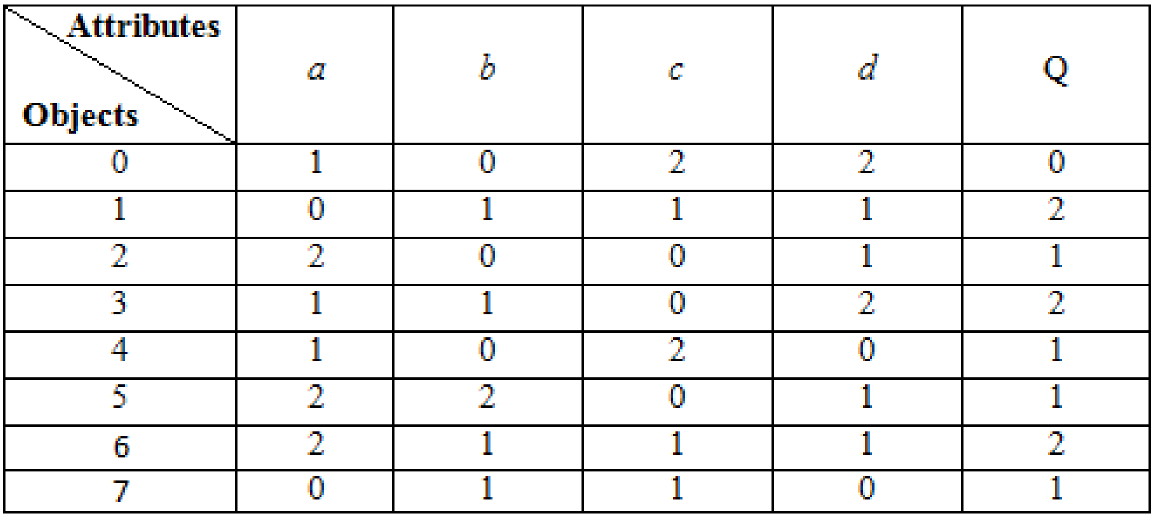
$(U, C \cup D, V, f)$ is called consistent if $\gamma_C(Q) = 1$.

Definition 1.9.1 *A subset P of C is called a positive region based decision reduct if it* *satisfies $POS_P = POS_C$, i.e., P preserves the decision making power of C, and moreover, it cannot be further reduced, i.e., there exists no proper subset P_0 of P such that $POS_{P_0} = POS_C$. If the latter constraint is lifted, i.e., P is not necessarily minimal, we call P a decision super reduct.*

Definition 1.9.2 *A subset P of C is called the degree of dependency based reduct if it satisfies $\gamma_P = \gamma_C$ with $\gamma_P \backslash \{a\} \neq \gamma_P, \forall a \in P$.*

Let us consider the following example dataset:

Table 1.5: Example of a Decision System



Attributes Objects	*a*	*b*	*c*	*d*	Q
0	1	0	2	2	0
1	0	1	1	1	2
2	2	0	0	1	1
3	1	1	0	2	2
4	1	0	2	0	1
5	2	2	0	1	1
6	2	1	1	1	2
7	0	1	1	0	1

For the illustrative example dataset, if $P = \{b, c\}$, then instances 1,6, and 7 are indiscernible from the instances 0 and 4. IND(P) generates the following partition of U :

$$U/(IND(P)) = U/(IND(b)) \bigotimes U/(IND(c))$$

where

$$A \bigotimes B = \{X \cap Y : \forall X \in A, \forall Y \in B, X \cap Y \neq \phi\}$$

$$= \{\{0, 2, 4\}, \{1, 3, 6, 7\}, \{5\}\} \bigotimes \{\{2, 3, 5\}, \{1, 6, 7\}, \{0, 4\}\}$$

$$= \{\{2\}, \{0, 4\}, \{3\}, \{1, 6, 7\}, \{5\}\}$$

$$POS_P(Q) = \bigcup\{\phi, \{2, 5\}, \{3\} = \{2, 3, 5\}$$

Now degree of dependency of Q over P can be given by

$\gamma_P(Q) = \frac{|POS_P(Q)|}{|U|} = \frac{|\{2,3,5\}|}{8} = \frac{3}{8}$.

From the example dataset, the dependencies for all possible subsets of conditional attributes can be calculated as follows:

$\gamma_{\{a\}}(Q) = \frac{0}{8}, \gamma_{\{b\}}(Q) = \frac{1}{8}, \gamma_{\{c\}}(Q) = \frac{0}{8}, \gamma_{\{d\}}(Q) = \frac{2}{8}$.

$\gamma_{\{a,b\}}(Q) = \frac{4}{8}, \gamma_{\{a,c\}}(Q) = \frac{4}{8}, \gamma_{\{a,d\}}(Q) = \frac{3}{8}, \gamma_{\{b,c\}}(Q) = \frac{3}{8}, \gamma_{\{b,d\}}(Q) = \frac{8}{8}, \gamma_{\{c,d\}}(Q) = \frac{8}{8}. \gamma_{\{a,b,c\}}(Q) = \frac{4}{8}, \gamma_{\{a,b,d\}}(Q) = \frac{8}{8}, \gamma_{\{a,c,d\}}(Q) = \frac{8}{8}, \gamma_{\{b,c.d\}}(Q) = \frac{8}{8}$.

$\gamma_{\{a,b,c,d\}}(Q) = \frac{8}{8}$.

It is obvious that the given dataset is consistent as $\gamma_{\{a,b,c,d\}}(Q) = 1$. The minimal reduct set for this example is

$R_{min} = \{\{b,d\},\{c,d\}\}$

So the core(intersection of all the possible reduct sets) of this dataset is $\{d\}$. Rough set based approach is one of the most important techniques of attribute selection that acquires information from the data set itself. Rough set theory does not need any external information for attribute selection.It handles the vagueness in the information system. However,this method can be applied to discrete data only. Therefore, discretization methods are applied in order to tackle the real-valued information system before attribute selection and this may lead to loss of some information.

In order to cope with this problem, fuzzy rough set (proposed by Dubois and Prade [20; 21]) based approach is presented to resolve both uncertainty and vagueness available in the data set. Combining Zadeh's fuzzy set (Zadeh, 1965) and rough set gives a key route in reasoning with uncertainty for real-valued data. Fuzzy rough set concept has been implemented to surpass the deficiencies of the classical rough set approaches.

1.10 Fuzzy Set Theory

Fuzzy set theory [49; 100] associates a membership value, i.e., the degree to which an element belongs to a set, to all the elements of a set. A fuzzy set in U is a mapping

$\mu : U \longrightarrow [0, 1]$ and $\mu_A(x)$ represents membership grade of x in A. A fuzzy relation in U is a fuzzy set defined on $U \times U$. For all $y \in X$, the fuzzy set R_y is the R-foreset of y, defined by

$$R_y(x) = R(x, y) \tag{1.6}$$

for all x in U. If R is a reflexive fuzzy relation, that is,

$$R(x, x) = 1 \tag{1.7}$$

and R is a symmetric fuzzy relation, that is,

$$R(x, y) = R(y, x) \tag{1.8}$$

hold $\forall x, y \in X$, then R is called a fuzzy tolerance relation, and R_y is referred to as the fuzzy tolerance class of y. For two fuzzy sets A_1 and A_2 in X :

(1). $A_1 \subseteq A_2 \Leftrightarrow (\forall x \in X)(\mu_{A_1}(x) \leq \mu_{A_2}(x))$.

(2). $\mu_{A_1 \cup A_2}(x) = sup\{\mu_{A_1}(x), \mu_{A_2}(x)\}$

(3). $\mu_{A_1 \cap A_2}(x) = inf\{\mu_{A_1}(x), \mu_{A_2}(x)\}$

(4). $\mu_{A_1{}^c}(x) = 1 - \mu_{A_1}(x)$

(5). If X is finite, the cardinality of the fuzzy set A_1, is calculated as:

$$|A_1| = \Sigma_{x \in X} \mu_{A_1}(x) \tag{1.9}$$

Definition 1.10.1 ***Fuzzy Triangular-norm:*** *[42] A triangular norm or t-norm, T is an increasing, associative and commutative mapping from* $[0, 1] \times [0, 1] \rightarrow [0, 1]$ *satisfying* $T(1, x) = x, \forall x \in [0, 1]$. *A few widely used t-norms are :* $T_M(x, y) = min\{x, y\}$ *and* $T_L(x, y) = max\{0, x + y - 1\}$ *(Lukasiewicz t-norm), for x, y in [0, 1].*

Definition 1.10.2 ***Fuzzy Implicator:*** *[42] An implicator is any* $[0, 1] \times [0, 1] \rightarrow [0, 1]$ *mapping* I, *satisfying* $I(0, 0) = 1$ *and* $I(1, x) = x, \forall x \in [0, 1]$. *Moreover, I needs to* *be decreasing in its first, and increasing in its second component. A few widely used implicators are-* $I_M(x,y) = max\{1-x,y\}$ *(Kleene- Dienes implicator) and* $I_L(x,y) = min\{1, 1-x+y\}$ *(Lukasiewicz implicator)* $\forall x, y \in [0,1]$.

1.11 Fuzzy-Rough Set Theory

The Fuzzy Rough Set Theory(FRST) [20; 21] is an alternate method for feature selection. It proposes to calculate the similarity between the objects using a fuzzy relation R in U, i.e., a $U \times U \rightarrow [0,1]$ mapping that assigns to each distinct pair of objects their corresponding degree of similarity, where R is a fuzzy tolerance relation, i.e., it follows Eq.(1.7) and Eq.(1.8). Given a set $X \subseteq U$ and a fuzzy tolerance relation R, the lower and upper approximation of X by R can be calculated in several ways. A general definition is the following:

$$(R \downarrow X)(x) = inf_{y \in U} I(R(x,y), X(y)) \tag{1.10}$$

$$(R \uparrow X)(x) = sup_{y \in U} T(R(x,y), X(y)) \tag{1.11}$$

Here, I is a fuzzy implicator and T is a fuzzy t-norm and $X(y) = 1$, for $y \in X$, Otherwise $X(y) = 0$ The FRS approximations defined in Eq.(1.10) and Eq.(1.11) are highly sensitive to noisy values, due to the use of sup and inf (the generalized existential and universal quantifier, respectively). It adversely affects the accuracy in case of classification problems.

1.12 Fuzzy Rough Set based Feature Selection

The rough set based feature selection works effectively only in datasets with discrete values, whereas most data in day to day life contain real values. Hence, the fuzzy rough set based feature selection was introduced which takes into consideration the similarity between the values of objects for a particular attribute. It helps in modelling data with the uncertainty involved in it.

Equivalence classes in the rough set based feature selection can be extended to the fuzzy rough based feature selection [39; 40; 41; 42; 43; 53; 80; 81; 85; 72; 88; 96] by introducing a fuzzy similarity relation, R in U, which determines degree of similarity of two objects on R [63]. The lower and upper approximations have been defined in Eq.(1.10) and Eq.(1.11) respectively.

Definition 1.12.1 ***Fuzzy Decision System:*** *A quadruple (U, A, V_F, F) is said to be a fuzzy information system, where U (the universe of discourse) is a non-empty set of finite objects, A is a non-empty finite set of attributes, V_F is the set of attribute values and F is an information function from $U \rightarrow V_F$. If $A = C \cup D$ such that $C \cap D = \phi$, where C is the set of conditional attributes and D is the set of decision attributes, then (U, A, V_F, F) is known as fuzzy decision system.*

Now, positive region can be defined by

$$\mu_{POS_P(Q)} = sup_{X \in U/Q} \mu_{R_P \downarrow X} \tag{1.12}$$

Here, $P \subseteq C$, $X \subseteq U$ and Q is the set of decision attributes.

Now, degree of dependency can be given by

$$\Upsilon_P(Q) = \frac{\mid \mu_{POS_P(Q)}(x) \mid}{\mid U \mid} = \frac{\Sigma_{x \in U} \mu_{POS_P(Q)}(x)}{\mid U \mid} \tag{1.13}$$

If the fuzzy-rough reduction process is to be useful, it must be able to deal with multiple features, finding the dependency between various subsets of the original feature set. For example, it may be necessary to determine the degree of dependency of the decision feature(s) with respect to $P = \{a, b\}$. In the crisp case, U/P contains sets of objects grouped together that are indiscernible according to both features a and b. In the fuzzy case, objects may belong to many equivalence classes, so the Cartesian product of $U/IND(a)$ and $U/IND(b)$ must be considered in determining U/P. In general,

$$U/P = \bigotimes\{a \in P : U/IND(a)\} \tag{1.14}$$

Each set in U/P denotes an equivalence class. For example, if $P = \{a, b\}, U/IND(a) = \{N_a, Z_a\}$ and $U/IND(b) = \{N_b, Z_b\}$, then $U/P = \{N_a \cap N_b, N_a \cap Z_b, Z_a \cap N_b, Z_a \cap Z_b\}$ The extent to which an object belongs to such an equivalence class is therefore calculated by using the conjunction of constituent fuzzy equivalence classes, say F_i, where $i = 1, 2, ..., n : \mu_{F_1 \cap ... \cap F_n}(x) = min\{\mu_{F_1}(x), \mu_{F_2}(x), ..., \mu_{F_n}(x)\}$.
A problem may arise when this approach is compared to the crisp approach. In conventional RSAR, a reduct is defined as a subset *Red* of the features which have the same information content as the full feature set A. In terms of the dependency function this means that the values $\gamma_{(Red)}$ and $\gamma_{(A)}$ are identical and equal to 1, if the dataset is consistent. However, in the fuzzy-rough approach this is not necessarily the case as the uncertainty encountered when objects belong to many fuzzy equivalence classes results in a reduced total dependency. The FRFS algorithm is given as follows [39] :

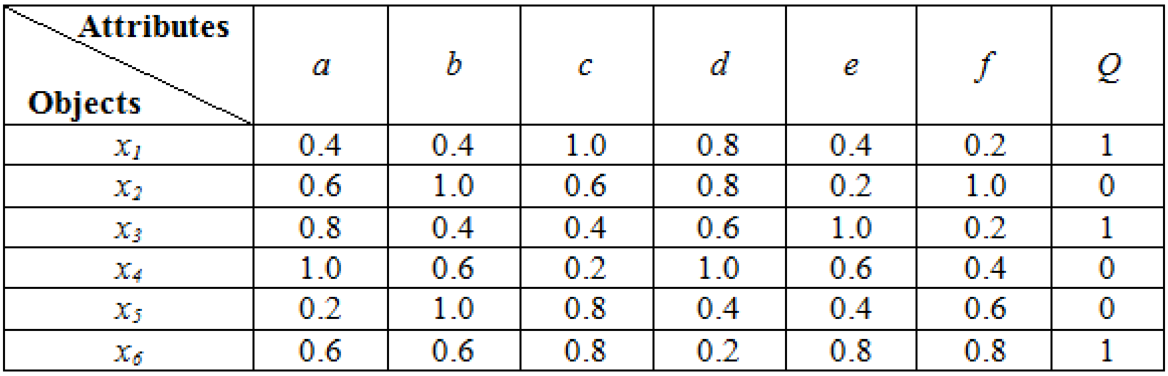
Fuzzy Rough Quick Reduct Algorithm (C,D)

```
C, the set of all conditional attributes;
D, the set of decision attributes.
$Red \leftarrow \{\}; \gamma_{best} = 0; \gamma_{prev} = 0$
do
$T \leftarrow Red$
$\gamma_{prev} = \gamma_{best}$
for each $x \epsilon (C \backslash Red)$
if $(\gamma_{Red \cup \{x\}})(D) > (\gamma_T)(D)$
$T \leftarrow Red \cup \{x\}$
$\gamma_{best} = (\gamma_T)(D)$
$Red \leftarrow T$
until $\gamma_{best} == \gamma_{prev}$
```

return *Red*

In order to illustrate our approach of fuzzy rough set based feature selection, a data set inspired from [42] is given in Table 1.6.

Table 1.6: Fuzzy Decision System



Attributes Objects	*a*	*b*	*c*	*d*	*e*	*f*	*Q*
x_1	0.4	0.4	1.0	0.8	0.4	0.2	1
x_2	0.6	1.0	0.6	0.8	0.2	1.0	0
x_3	0.8	0.4	0.4	0.6	1.0	0.2	1
x_4	1.0	0.6	0.2	1.0	0.6	0.4	0
x_5	0.2	1.0	0.8	0.4	0.4	0.6	0
x_6	0.6	0.6	0.8	0.2	0.8	0.8	1

From Table 1.6, decision class can be given as follows:

$U \backslash Q = \{(x_1, x_3, x_6), (x_2, x_4, x_5)\}$ Degree of dependencies of Q over $A = \{a\}, B = \{b\}, C = \{c\}, D = \{d\}, E = \{e\}, F = \{f\}$ can be calculated using [31] as follows:

$\gamma_A(Q) = \frac{1.2}{6}, \gamma_B(Q) = \frac{2.4}{6}, \gamma_C(Q) = \frac{1.2}{6}, \gamma_D(Q) = \frac{1.2}{6}, \gamma_E(Q) = \frac{2.2}{6} and \gamma_F(Q) = \frac{1.2}{6}$ Since feature $\{b\}$ will cause the greatest increase in dependency degree. Hence, this feature is chosen and added to the potential reduct set.

Now adding other features to potential reduct set, we calculate degree of dependencies for $\{a, b\}, \{b, c\}, \{b, d\}, \{b, e\}, \{b, f\}$ as:

$$\gamma_{\{a,b\}}(Q) = \frac{2.2}{6}, \gamma_{\{b,c\}}(Q) = \frac{2.2}{6}, \gamma_{\{b,d\}}(Q) = \frac{2.6}{6}, \gamma_{\{b,e\}}(Q) = \frac{2.2}{6}, and \gamma_{\{b,f\}}(Q) = \frac{2.0}{6}.$$

On adding feature $\{d\}$ to the reduct candidate causes the larger increase of degree of dependency. So, new reduct becomes $\{b, d\}$ This process iterates, and we get other degrees of dependencies as

$\gamma_{\{a,b,d\}}(Q) = \frac{2.4}{6}, \gamma_{\{b,c,d\}}(Q) = \frac{2.2}{6}, \gamma_{\{b,d,e\}}(Q) = \frac{2.2}{6}, and \gamma_{\{b,d,f\}}(Q) = \frac{2.2}{6}$

It is obvious that by adding rest of the features with cause no increase in degree of dependency, the algorithm stops and outputs the reduct $\{b, d\}$.

1.13 Limitations of Fuzzy Set

Fuzzy set theory is a powerful tool to deal with uncertainty but it has certain limitations also. These limitations can be outlined as follows:

- Fuzzy set theory is not capable of handling many decision making problems, for example, in a voting problem, where, a panel of 10 experts are voting for 1 item, suppose 5 experts gave the conclusions "agree", three of them "opponent" and rest of the two experts "abstain". This scenario can be effectively handled by adding a non-membership degree for "opponent" and hesitancy degree for "abstain".
- In the many real world applications, such as medical diagnosis and sensor information etc., vaguely specified data values are very common. Fuzzy set theory is implemented to handle such vagueness by generalizing the concept of membership in a set. In a fuzzy set, the membership degree of the element in a universe always takes a single value between 0 and 1 but those single values may not completely define about the lack of knowledge as the uncertainty is not found only in judgement but also in the identification. Therefore, some extensions of fuzzy sets are required to handle the latter uncertainty.

In order to tackle more uncertain and complex information system, many concepts have been presented. A vague set and intuitionistic fuzzy set [1; 2; 3] are two well-known extensions of fuzzy set. In a vague set, interval based membership is used. Interval-based membership is more expressive in capturing vagueness available in data. Bustince and Burillo (1996) [6] investigated about vague set and intuitionistic fuzzy set and showed that they are equivalent. Intuitionistic fuzzy set based approaches are widely used to handle uncertainty, in which membership, non-membership and hesitancy functions are considered simultaneously. Thereby, it can handle uncertainty in much better way when compared to fuzzy approaches. So, it has much stronger ability to deal with information system and draw a better glimpse of fragile ambiguities of the objective world [70].

Intuitionistic fuzzy set is the suitable choice in the situation when representation of non-membership degree is found to be simpler than membership degree. Therefore, it is anticipated that the human decision-making process and activities requiring human expertise and knowledge which are inevitably imprecise or not totally reliable could be used to simulate by using intuitionistic fuzzy set concept. Intuitionistic fuzzy set based approach has been successfully implemented in decision making concept and pattern recognition. The key benefits of intuitionistic fuzzy sets over fuzzy sets are:

- A vague pattern classification can be transformed into a precise and well defined optimization problem by using intuitionistic fuzzy set approaches.
- Unlike fuzzy sets, intuitionistic fuzzy sets preserve a precise degree of the uncertainty.

1.14 Intuitionistic Fuzzy Set

Definition 1.14.1 *An ordered pair $\langle m, n\rangle$ is said to be an intuitionistic fuzzy value if $0 \leq m, n \leq 1$ and $0 \leq m + n \leq 1$. Let U be an universe of discourse, which represents finite collection of objects, and L is an intuitionistic fuzzy set in U such that*

$$L = \{\langle x, m_L(x), n_L(x)\rangle | x \in U\} \tag{1.15}$$

where, $m_L \colon U \to [0, 1]$ and $n_L \colon U \to [0, 1]$ satisfy $0 \leq m_L(x) + n_L(x) \leq 1$, $\forall x \in U$, then $m_L(x)$ and $n_L(x)$ are respectively known as degree of membership and degree of non-membership of the element $x \in U$ to L.

Definition 1.14.2 *Let $\langle m_i, n_i\rangle$ be two intuitionistic fuzzy values for $i = 1, 2$, then, following properties hold,*

(a) $\langle m_1, n_1\rangle = \langle m_2, n_2\rangle \Leftrightarrow m_1 = m_2 \wedge n_1 = n_2$

(b) $\langle m_1, n_1\rangle \cap \langle m_2, n_2\rangle = \langle min\{m_1, m_2\}, max\{n_1, n_2\}\rangle$

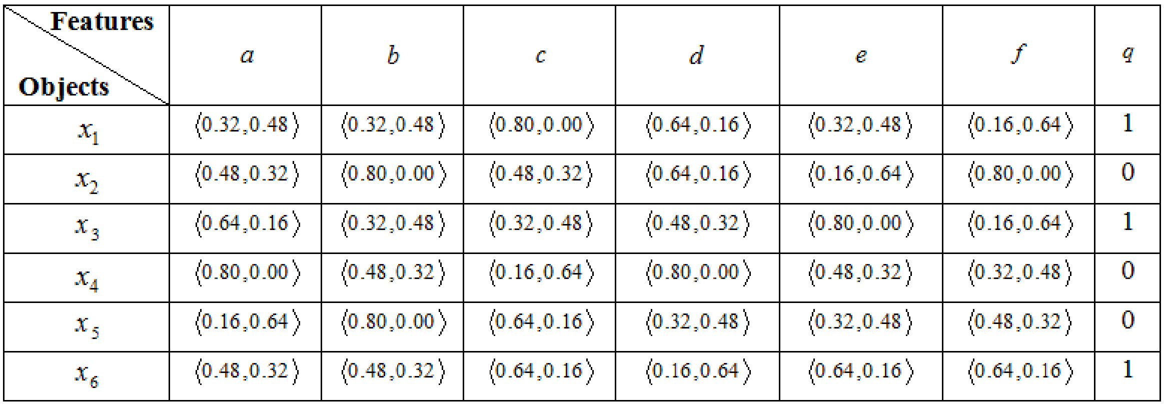
(*c*) $\langle m_1, n_1\rangle \cup \langle m_2, n_2\rangle = \langle max\{m_1, m_2\}, min\{n_1, n_2\}\rangle$

(*d*) *The cardinality of intuitionistic fuzzy set L is defined by*

$$|L| = \Sigma_{x\in U}\frac{1 + m_L(x) - n_L(x)}{2} \tag{1.16}$$

Definition 1.14.3 *[31; 32; 33; 34] An intuitionistic fuzzy decision system can be defined as a quadruple* $IFIS = (U, C \cup D, V_{IF}, IF)$ *where* $U(\neq \phi)$ *is collection of finite number of objects, called universe of discourse,* $C(\neq \phi)$ *and* D *are finite sets of conditional and decision features such that* $C \cap D = \phi$ *,* V_{IF} *is the collection of all intuitionistic fuzzy values such that* $V_{IF} = S_1 \cup S_2$ *, where* S_1 *and* S_2 *are domains of conditional and decision features and* IF *is called information function which is defined as* $IF\colon U \times C \cup D \to V_{IF}$*, such that* $IF(x, a) \in S_a, \forall a \in C, S_a \subseteq S_1$ *and* $IF(x, Q) \in S_2$ *for* $D = \{Q\}$*, where,* $IF(x, a)$ *and* $IF(x, Q)$ *are intuitionistic fuzzy values. Table 1.7 represents an example of intuitionistic fuzzy decision system.*

Table 1.7: Intuitionistic Fuzzy Decision System



Features Objects	*a*	*b*	*c*	*d*	*e*	*f*	*q*
x_1	⟨0.32,0.48⟩	⟨0.32,0.48⟩	⟨0.80,0.00⟩	⟨0.64,0.16⟩	⟨0.32,0.48⟩	⟨0.16,0.64⟩	1
x_2	⟨0.48,0.32⟩	⟨0.80,0.00⟩	⟨0.48,0.32⟩	⟨0.64,0.16⟩	⟨0.16,0.64⟩	⟨0.80,0.00⟩	0
x_3	⟨0.64,0.16⟩	⟨0.32,0.48⟩	⟨0.32,0.48⟩	⟨0.48,0.32⟩	⟨0.80,0.00⟩	⟨0.16,0.64⟩	1
x_4	⟨0.80,0.00⟩	⟨0.48,0.32⟩	⟨0.16,0.64⟩	⟨0.80,0.00⟩	⟨0.48,0.32⟩	⟨0.32,0.48⟩	0
x_5	⟨0.16,0.64⟩	⟨0.80,0.00⟩	⟨0.64,0.16⟩	⟨0.32,0.48⟩	⟨0.32,0.48⟩	⟨0.48,0.32⟩	0
x_6	⟨0.48,0.32⟩	⟨0.48,0.32⟩	⟨0.64,0.16⟩	⟨0.16,0.64⟩	⟨0.64,0.16⟩	⟨0.64,0.16⟩	1

Definition 1.14.4 *[15]Let* U *be collection of finite objects, then an intuitionistic fuzzy binary relation* R *is called an intuitionistic fuzzy equivalence relation if and only if it is reflexive* ($\mu_R(x, x) = 1$ *and* $\nu_R(x, x) = 0$)*, symmetric* ($\mu_R(x, y) = \mu_R(y, x)$ *and* $\nu_R(x, y) = \nu_R(y, x), \forall x, y \in U$) *and transitive intuitionistic fuzzy relation* ($\mu_R(x, z)$ $\geq \vee(\mu_R(x, y) \wedge \mu_R(y, z))and(\nu_R(x, z) \leq \wedge(\nu_R(x, y) \vee \nu_R(y, z))), \forall x, y, z \in U$)*, An intuitionistic fuzzy binary relation* R *is called an intuitionistic fuzzy tolerance relation if it is reflexive and symmetric relation on* U*.*

Definition 1.14.5 *[13] Let* $(U, C \cup D, V_{IF}, IF)$ *be an intuitionistic fuzzy decision system.* [5, 90] *Let* T_1 *denotes intuitionistic fuzzy triangular norm,* I_1 *is an intuitionistic fuzzy implicator and* R *represents intuitionistic fuzzy equivalence relation in* U*. Let* X *be any intuitionistic fuzzy set in* U*, then the lower and upper approximations of* X *can be given by*

$$(R \downarrow X)(x) = inf_{y \in U} I_1(R(x, y), X(y)), \tag{1.17}$$

$$(R \uparrow X)(x) = sup_{y \in U} T_1(R(x, y), X(y)), \tag{1.18}$$

A couple of intuitionistic fuzzy sets $\{X_1, X_2\}$ *is called an intuitionistic fuzzy rough set in the* $(U, J \cup K, S, F)$ *information system, if there exists an intuitionistic fuzzy set* X *, such that* $R \downarrow X = X_1$ *and* $R \uparrow X = X_2$*.*

1.15 Literature Survey

In spite of the fact that rough sets and intuitionistic fuzzy sets both capture specific aspects of the same idea-imprecision, the combination of intuitionistic fuzzy set theory and rough set theory [60] are rarely discussed by the researchers. Jena, Ghosh, and Tripathy (2002)[38] ,Nanda and Majumdar (1992)[65] and Chakrabarty, Gedeon, and Koczy (1998) [8] demonstrated that lower and upper approximations of intuitionistic fuzzy rough sets are again intuitionistic fuzzy sets. Samanta and Mondal (2001)[74] presented a similar idea. Lu and Lei (2009)[61] presented a novel intuitionistic fuzzy rough set model by using distance concept to resolve many real-life conflict problems. In the last few years, some of the intuitionistic fuzzy rough set models have been established by many researchers established relationship between rough set and intuitionistic fuzzy set and revealed the fact that fuzzy rough set is admittedly an intuitionistic L-fuzzy set [12] . Nowadays, the intuitionistic fuzzy rough set theory is emerging as an effective and powerful tool to deal with uncertainty and applied for decision making to solve many real life problems [7, 9, 10, 16, 73]. However, very few research works have been introduced in the area of feature selection based on intuitionistic fuzzy rough set.LU, LEI, and HUA (2009)[61] proposed the genetic algorithm for attribute reduction of IFIS. Chen and Yang (2011)[11] combined intuitionistic fuzzy rough set with information entropy and introduced a new attribute reduction algorithm. Esmail, Maryam, and Habibolla (2013)[23] studied about the structure of intuitionistic fuzzy rough set model and its properties and presented their method of attribute reduction along with rule extraction. Huang, Li, and Wei (2012)[**?**] designed an intuitionistic fuzzy rough set based attribute reduction model by using distance function. Z. Zhang (2016)[106] presented an attribute reduction method by using discernibility matrix approach. However, none of the proposed approaches for feature selection is based on dependency function by considering similarity between two objects.

The concept of a dependency function in a traditional rough set model into the fuzzy occurrence was proposed by Jensen and Shen (2004)[39] and introduced a feature selection algorithm using fuzzy rough set concept. This concept was extended by many researchers in their research articles [39; 40; 41; 42; 43; 53; 80; 81; 85; 72; 88; 96]. Dependency function based approaches perform better than discernibility matrix based approaches in fuzzy case (Jensen and Shen(2008)[42]). In the current thesis, dependency function based approaches for feature selection using intuitionistic fuzzy rough set models have been presented. These approaches are applied on arbitrary datasets.

****